CONTROLLING FOR HIDDEN FACTORS IN HIGH DIMENSIONAL eQTL STUDIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Chuan Gao

May 2012

CONTROLLING FOR HIDDEN FACTORS IN HIGH DIMENSIONAL eQTL STUDIES

Chuan Gao, Ph.D.

Cornell University 2012

Finding genetic variants that regulate gene expression now plays a central role in the analysis of mechanism in biological systems. This will also increasingly be the case as large amounts of gene expression and genetic marker data are generated by next-generation sequencing technologies. While the unprecedented scale of these data is providing the opportunity for scientists to answer basic questions about biological systems, the properties of these data raise analysis challenges, particularly in terms of covariate modeling. For example, expression levels of thousands of genes are usually measured in batches and different batches may be measured under different conditions, which creates the well known batch effect. Besides this artificially created factor that can affect the quality of the measurement, expression data often reflect environmental regulators that change the gene expression levels, such as smoking, drug usage etc.. These sources of confounding need to be addressed either before or during analysis of data.

In this thesis, I address the analysis issues raised by a particular type of confounding in high-dimensional data: hidden factor effects. Hidden factors are defined as factors that contribute to variation in a large number of measured variables where there is no direct information concerning the factors in the data. It is critical to correct for the hidden factors because if ignored, they can lead to either high false positive rates or reduced power. To tackle this issue, I propose

to use a statistical model that combines multivariate ridge regression and factor analysis to infer both the fixed effects and the hidden confounding. The method is unique in the sense that it employs the multivariate regression components to infer the associations between the response Y and the covariate X, while it maintains efficiency by sharing the same data reduction property with the factor analysis model. Compared to other models that address the same issue, this model can successfully partition the covariance structure of the hidden factors, which dramatically improves the power and the accuracy of detecting the real associations between X and Y. I also used the model to address the hidden factors issues in the analysis of data on gene expression levels measured in the airway of the lung in a sample of people, in the context of a genome association study, referred to as an expression Quantitative Trait Loci (eQTL) analysis. I show that the method successfully eliminates the false positives caused by spurious structures (hidden factors) and greatly improves the power to detect true genetic determinants (the eQTL) that regulate gene expression in the lung airway. I also apply the method to a challenging Genotype-Environment Interaction (GEI) analysis, where GEI effects are defined as the dependence of genotype-phenotype relationships on environmental factors. I show that despite the small sample size and the highly complicated data structure, with my method, I can identify a large number of interesting GEI associations, many have been verified indepently by other studies to be highly relevant genes to lung disease and lung functions. These GEI associations contain more information than a typical eQTL because they help to identify genetic regulators that show different behavior under different environmental pressures, which serve as an interesting set of gene candidates for clinical scientists.

BIOGRAPHICAL SKETCH

Chuan Gao obtained his bachelor and master degree respectively from the microbiology and physiology program in the department of biology at China Agricultural University. Then he went to Texas A&M University to study in the program of animal breeding, where his major work was to find genetic determinants that help to improve the meat quality of the livestock. He was fascinated by the statistics behind the analysis and decided to apply for the computational biology program at Cornell University, where he was luckily admitted and supervised by Dr. Jason G. Mezey. For his six years of Ph.D study in Cornell, he devoted most of the time to developing methods for Genome Wide Association Studies (GWAS) and high dimensional expression Quantitative Trait Loci (eQTL) analysis, trying to identify interesting novel QTLs and eQTLs. He discovered the dynamics of the fixed and random effects in the random effect model and based on this, he developed a statistical method, Hidden Expression FacTor analysis (HEFT), to address the hidden factors problem in eQTL analysis. In his spare time, he enjoyed analyzing all sorts of disease and gene expression data and making interesting discoveries.

To my family

ACKNOWLEDGEMENTS

Completing my Ph.D study is one of the most important things I have accomplished in my life, and I want to share this accomplishment with whom I have been working with and who have showed me support.

My biggest thanks goes to my Ph.D advisor, Dr. Jason G. Mezey. It is difficult to overstate his kindness, passion and inspiration. Jason took me as his student during my most challenging time when I was learning statistics as a biology major, and has been making statistics a fun subject with his clear explanations and patient guidance. During my six years of Ph.D. study, he provided me encouragement when I was frustrated, and sound advice on both my thesis writing and career. This thesis would never have been possible without him.

Special thanks to my committee members Dr. Andrew (Andy) G. Clark, Dr. James (Jim) Booth and Dr. Ping Li for their guidance and helpful suggestions. Andy was the founder and the chair of the Computational Biology program back at the time when I entered Cornell and his name was always connected to my early time in the program. I also owe him gratitude for finding me a three year Provost fellowship that supported me through half of my Ph.D. career, and his deep insight into the -omics world and broader genetics and biological problems. Jim is the chair of the department of Biological Statistics and Computational Biology and, despite his busy schedule, he was always available for help, spending time deriving formulas with me on the blackboard, which greatly boosted my confidence on problems that I was less certain about. I also learned a lot from Ping, who is both a computational statistician and a machine learning scientist. Ping's suggestions and discussions on my machine learning model and his computational skills have been a great benefit to me.

I would like to thank my colleagues, Benjamin Logsdon, Gabriel Hoffman,

Lin Li, Keyan Zhao, Hong Gao, Fangfei Ye, Pavel Korniliev, Anthony Greenberg, Yuxin Shi, Francisco Agosto-Perez, Larsson Omberg, Haley Hunter-Zinck, Cristopher V. Van Hout for their helpful discussions and delightful company. Working with them is full of fun and pleasure.

I also want to thank my parents and my grandparents for their selfless love, and my wife Yan Zhao, my son Ethan Gao who made my life more joyful and meaningful. Special thanks to my brother Roy and sister Yuan, who helped me financially through my undergraduate study and pointed me in the right direction. Knowing that they are standing behind me makes me strong.

There are also many others too numerous to list that have made contributions to me during my studies and my sincere thanks goes to them as well.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTIONS

The development of second generation sequencing technologies has produced an overwhelming amount of genetic and genomic data [1, 2, 3], which has opened opportunities for biologists to use these data to answer questions about biological systems. For example, using the millions of Single Nucleotide Polymorphisms (SNPs) from the hapmap project [4], combined with the case control data collected by medical facilities, many genetic determinants for a variety of disease have been mapped onto a small region of the genome by using the Genome-Wide Association Study technique (GWAS)[5], shedding light on the genetic causes of these diseases. These genetic analyses can also borrow strength from the analysis of the large array of gene expression data obtained using either microarray technology or from the more recent RNA-Seq technology, yielding more accurate results[6]. These latter data can also help elucidate the disease mechanism by interrogating the behavior of the transcripts that are associated with the disease, using the expression Quantitative Trait Loci analysis method (eQTL) [7].

Like GWAS, where a phenotype of interest is analyzed by scanning through the whole genome for significant associations with genetic markers such as SNPs, eQTL analysis does the same, except that thousands of genes are taken as the phenotypes. eQTL analysis is becoming more and more important because it plays a central role in connecting the genetic determinants of the diseases to critical causal genes, and its output is critical for downstream analyses, e.g. the informative eQTLs and genes can be used to dissect gene pathways [8, 9] or for gene network analysis [10, 11], providing further insight into the mechanisms

of a disease [7]. Despite the rapid progress made by GWAS and eQTL analysis, problems remain for both methods when detecting genetic variantes in the relatively low heritability range [12, 13, 14], where heritability measures the fraction of phenotype variability that can be attributed to the additive effects of genetic variation. Most of the discoveries made by these two methods employed the one genotype at a time Simple Linear Regression testing technique, treating the phenotype of interest as y and the single genotype as x. However, this simple model structure can cause problems when underlying covariates exist. A well known problem for GWAS is population structure, where different ethnicity backgrounds within the sample can cause large amounts of false positives. Extensive effort has been put into addressing this problem which has resulted in dramatic improvements [15, 16, 17, 18, 19, 20], but room remains for further improvement.

The covariate problem is exacerbated in eQTL analysis which can result in even larger numbers of false positives or power reduction for detecting real genetic associations. Since the gene expression levels are typically measured on a large scale to characterize the transcription abundance genome-wide, problems that exist in the GWAS analysis can manifest themselves in eQTL analysis at a higher magnitude. As a result, finding the real association among the large number of false positives poses a serious challenge, let alone the fact that low power may prevent the real associations entering the top associations list. These problems are typically caused by spurious covariates with structures that are either correlated or uncorrelated with the genotypes. For the former, this includes the population structure problem, and methods for correcting these structures in GWAS analysis have been used in the analysis of eQTLs. For the later, this in-

cludes confounding factors that can be independent of the genotypic effects. These structures are usually produced by environmental effects such as batch effects from the microarray experiment that bias the raw data, or environmental factors like smoking [21], alcohol use [22], pressure [23] etc., that can affect the gene expression level on a large scale. Ideally, information on these factors should be collected, which can then be incorporated into the analysis later as known covariates. The problem is not every confounding factor can be envisioned in advance, so no matter how well the experimental design is controlled, there is no guarantee that all confounding sources have been considered. There is therefore a need to post-process the data so that these unaccounted for hidden confounding factors can be addressed.

Ideally, hidden confounding structures should be modeled directly in a statistical analysis and several methods have been proposed to accomplish this [24, 25, 26, 27, 28, 29, 30, 31], each with their own advantages and limitations that will be explained in more detail in the later chapters. In this work, I have addressed the hidden factor issue in eQTL analysis by using a very distinct Regression-Factor analysis model. I noticed some interesting properties about the model and by focusing on these properties, I was able to construct a likelihood based Ridge regression and factor analysis statistical model that can simultaneously control for hidden factors while inferring the genotypic effects that have effects on a large number of genes. I showed by extensive simulations that this method outperforms other competing methods in both improving the power of detecting genetic associations in eQTL analysis and in reducing the number of false positives. I also applied the method to a real data set comprising thousands of genes and hundreds of thousands of SNPs from the Small

Airway Epithelium (SAE) in the lung and discovered a large number of interesting eQTLs that are lung disease related, among which a non-trivial number have been confirmed independently by previous studies.

The inherent property of increasing the power and reducing false positives of my method make it appropriate for some of the most challenging statistical genomics problems, for example, in low powered Genotype-Environment Interaction (GEI) analyses. GEI describes the difference in phenotype in response to the effects of genotypic determinants under different environmental conditions. GEI is becoming more and more important because it reveals informative genetic information as to how the genes responded to different environments. However, it is well known that GEI analysis is low powered, largely because a bigger sample size is needed for GEI to yield comparable sample size for each genotype-environment category as compared to a simple QTL analysis, and the power can be further reduced by the existence of other complicated hidden factors. I applied my hidden factor method to the same SAE data set in the lung, where the samples includes both smokers and non-smokers, which can have effects on eQTL. The GEI analysis in these data is particularly interesting because it reveals genetic variants that regulate genes differently under smoking pressure, and because smoking is a well known cause of a variety of lung diseases, these GEI associations are highly informative candidates that clinical scientists should focus on. With my hidden factor analysis, I found a large number of interesting GEI candidates that are relevant for lung function or lung disease.

This thesis is structured in three main chapters. The second chapter focuses on an unrestricted and restricted Regression-Factor analysis model that com-

bines multivariate regression and factor analysis methods to infer both fixed effects and hidden factors. I reasoned that the unrestricted model can be used to address the hidden factors while inferring the fixed effects, however, only in a limited number of circumstances. By focusing on some of the most fundamental properties of the model, I then improve the model by imposing a constraint on the fixed effects, where the new restricted model: Hidden Expression FacTor analysis (HEFT) combines the multivariate ridge regression and the factor analysis. HEFT was able to address all hidden factor issues according to both my extensive simulations and in my application of the method to real data, where these analyses are described in more details in chapter 3. In the final chapter, I applied the HEFT model to the Genotype-Environment Interaction (GEI) problem, which typically requires a sample size of thousands to achieve reasonable power. I showed that by controlling for the hidden factors with HEFT, I can identify interesting novel GEI effects in these data. Throughout, I discuss the advantages and limitations of my statistical model in controlling for the hidden factors, and I propose further improvements for the model that could be incorporated in the future.

CHAPTER 2

DYNAMICS OF FIXED AND RANDOM EFFECTS, A RESTRICTED REGRESSION-FACTOR MODEL

In this chapter, I propose a statistical model that combines two methods, multivariate ridge regression and factor analysis to address the heterogeneity issue that is typically encountered in data sets with high dimension. The model was motivated by a high dimensional gene expression data set that contains measurements of hundreds of samples and tens of thousands of genes, where unobserved covariates, or latent factors can affect a large number of them, creating heterogeneity problems for further analysis of the data set. However, the model is not limited to these types of data, it can be applied to all sorts of high dimensional data where hidden structures are suspected to affect non-trivial number of the variables. Similar models have been proposed [32, 24, 26, 25, 27, 28, 29, 30, 31], which use distinct inference procedures for the parameters than the methods proposed here. These different treatments of the problem can lead to different performances when analyzing data.

Here, I give a detailed formulation of this model and describe how I use this model to simultaneously correct for heterogeneity. I first present the theory behind an iterative form of the algorithm assuming an unconstrained fixed effect, then I give the form of the algorithm for the constrained fixed effects as well the calculation of the p values. Underlying the theory of this work is a long ignored issue of fixed and random effects under a linear system with non-orthogonal fixed and random effects, where the term non-orthogonal refers to correlated fixed and random effects. Traditionally, people deal with this system with a linear mixed model where the random effects were integrated out, and the fixed

effects were inferred using the conditional likelihood. Instead, I use a factor analysis approach by explicitly modeling the random effects as observed variables. I explore the properties of this model and discuss the difference between this model and others. For all cases below, I work on a factor analysis model where a correlation matrix among the factors is used, where this will be useful when there are correlation structures among the hidden factor that can't be captured by an orthogonal assumption. However, in practice, I did notice that the orthogonal assumption is sufficient to capture the hidden factor, and I therefore only implemented a version of the algorithm with correlation matrix of identity. The same algorithm for the sole purpose of factor analysis without including the fixed effects has been presented by Rubin [33].

2.1 The model with unrestricted fixed effects

The full regression model with hidden structures can be written as

$$\mathbf{Y} = \mu\mathbf{1}^{'} + \mathbf{X}\beta + \Lambda\mathbf{F} + \mathbf{W} \tag{2.1}$$

where $\mathbf{Y}$ is the $n \times m$ response matrix, μ is the global parameter capturing the mean of each sample. $\mathbf{X}$ is the $n \times l$ fixed effect of interest with first column set to 1. $\mathbf{\Lambda}$ is the $n \times p$ loading matrix that load onto each sample, $\mathbf{F}$ is the $p \times m$ score of the hidden factor. and $\mathbf{W}$ is the $n \times m$ error matrix with $\mathbf{W} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Psi})$. I assume the first part of the model $\mathbf{X}\beta$ utilizes the multivariate regression approach, and the second part of the model $\mathbf{\Lambda F}$ utilizes the factor analysis approach, where $\mathbf{\Lambda F}$ should capture the correlation structure among the data matrix $\mathbf{Y}$, leading to correction of the heterogeneity of the data.

Generally, when people estimate the fixed effect β with confounding in a linear

mixed model framework, the incomplete likelihood is written as the following [20],

$$\mathbf{L}(\theta|\mathbf{D}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{nm/2}}|\mathbf{\Psi}+\mathbf{\Sigma}|^{\mathbf{m/2}}}\mathbf{exp}\left(\mathbf{tr}(-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{Y}-\mathbf{X}\beta)^{\mathbf{T}}(\mathbf{\Psi}+\mathbf{\Sigma})^{-\mathbf{1}}(\mathbf{Y}-\mathbf{X}\beta))\right) \tag{2.2}$$

where $\mathbf{\Sigma}$ is a similarity matrix obtained in advance to capture the covariance structure of the hidden factor, and the parameter estimator in the form of

$$\beta = (\mathbf{X}^{\mathbf{T}}(\mathbf{\Sigma}+\mathbf{\Psi})^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}(\mathbf{\Sigma}+\mathbf{\Psi})^{-\mathbf{1}}\mathbf{Y} \tag{2.3}$$

helps on correcting the spurious fixed effects caused by non-orthogonal structure, however, at a price of reducing power because the approach pools the error term and the covariance of hidden confounding instead of partitioning them. To see how the power can be reduced for this treatment of the model, we look at the variance term of β, which takes the form

$$(\mathbf{X}^{\mathbf{T}}(\mathbf{\Sigma}+\mathbf{\Psi})^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}} \tag{2.4}$$

where we see that the $Var(\beta)$ is confounded by the covariance matrix of the random effect $\mathbf{\Sigma}$, and this directly leads to deflation of the p values if we use a t test to calculate its p values. the same thing happens for a Likelihood Ratio Test (LRT), where the likelihood for both the full model and null model has been reduced, leading to a smaller $-2log\frac{L_{null}}{L_{full}}$.

I approach the problem from the factor analysis angle by explicitly modeling the hidden confounding with the goal of partitioning its variance from the error term, while simultaneously getting the estimate for the fixed effect. The complete likelihood with the hidden factors stated explicitly takes the following form

$$
\begin{aligned}
l_c(\theta_c|D_c) = \; & - \frac{\mathbf{pm}}{\mathbf{2}}\mathbf{log}(\frac{\mathbf{1}}{\mathbf{(2\pi)}}) - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\mathbf{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}(\mathbf{F}\mathbf{F}^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}) \\
& - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\mathbf{\Psi}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}((\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta - \mathbf{\Lambda}\mathbf{F})^{\mathbf{T}}\mathbf{\Psi}^{-\mathbf{1}}(\mathbf{y} - \mu\mathbf{1}' - \mathbf{X}\beta - \mathbf{\Lambda}\mathbf{F}))
\end{aligned}
\tag{2.5}
$$

Note that the term $\frac{pm}{2}log(\frac{1}{(2\pi)})$ is a function of the factor number, which can only be treated as a constant when a constant number of factors is involved. However, when different number of factors are involved, for example, in the situation where a different factor number needs to be selected based on some criteria involving the likelihood, this term has to be incorporated.

Usually, we work on the more convenient traces form by substituting the quadratic part by the following and use a matrix notation,

$$
\begin{aligned}
l_c(\theta_c|D_c) = \; & - \frac{\mathbf{pm}}{\mathbf{2}}\mathbf{log}(\frac{\mathbf{1}}{\mathbf{(2\pi)}}) - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\mathbf{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}(\mathbf{F}\mathbf{F}^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}) \\
& - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\mathbf{\Psi}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}((\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta - \mathbf{\Lambda}\mathbf{F})(\mathbf{y} - \mu\mathbf{1}' - \mathbf{X}\beta - \mathbf{\Lambda}\mathbf{F})^{\mathbf{T}}\mathbf{\Psi}^{-\mathbf{1}})
\end{aligned}
\tag{2.6}
$$

Next, I lay out the necessary pieces for the parameter inferences treating it as a factor analysis model.

2.2 The Expectation and Maximization Algorithm for the model

2.2.1 The Expectation step

In the expectation step, we transform the incomplete likelihood in equation 2.2 to the complete likelihood in equation 2.6, where this transformation is made

possible by noticing that the hidden factor $\mathbf{F}$ can be substituted by its expected value conditional on $\mathbf{Y}$. To get $\mathbf{E}(\mathbf{F}|\mathbf{Y})$, we note that the joint distribution of $\mathbf{F}$ and $\mathbf{Y}$ can be written as

$$\begin{pmatrix} \mathbf{F} \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \mu 1' + \mathbf{X}\beta \end{pmatrix}\begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Sigma\Lambda^T} \\ \mathbf{\Lambda\Sigma} & \mathbf{\Lambda\Sigma\Lambda^T} + \psi \end{pmatrix}\right) \tag{2.7}$$

from which the conditional variance of F can be written as

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma} - \mathbf{\Sigma\Lambda^T}(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^{-1}\mathbf{\Lambda\Sigma^T} \tag{2.8}$$

and the conditional expected value of $\mathbf{F}$ takes the the following form

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma\Lambda^T}(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^{-1}(\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta) \tag{2.9}$$

We can transform the high dimension expression $(\mathbf{\Lambda\Sigma\Lambda^T} + \mathbf{\Psi})^{-1}$ into a lower dimension form by utilizing the following transformation.

$$\mathbf{AB^T}(\mathbf{BAB^T} + \mathbf{R})^{-1} = (\mathbf{A^{-1}} + \mathbf{B^T R^{-1} B})^{-1}\mathbf{B^T R^{-1}} \tag{2.10}$$

and write $\mathbf{V}(\mathbf{F}|\mathbf{Y})$ as

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = \mathbf{\Sigma} - (\mathbf{\Sigma^{-1}} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}\mathbf{\Lambda^T\Psi^{-1}\Lambda\Sigma^T} \tag{2.11}$$

For the special case with $\mathbf{\Sigma} = \mathbf{I}$ We have

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = \mathbf{I} - (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}\mathbf{\Lambda^T\Psi^{-1}\Lambda} \tag{2.12}$$

we can further simplify the expression by the following

$$\begin{aligned} \mathbf{V}(\mathbf{F}|\mathbf{Y}) &= \mathbf{I} - (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}\mathbf{\Lambda^T\Psi^{-1}\Lambda} \\ &= (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}(\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda}) - (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}\mathbf{\Lambda^T\Psi^{-1}\Lambda} \\ &= (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1}(\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda} - \mathbf{\Lambda^T\Psi^{-1}\Lambda}) \\ &= (\mathbf{I} + \mathbf{\Lambda^T\Psi^{-1}\Lambda})^{-1} \end{aligned} \tag{2.13}$$

Using the same type of geometric trick

$$\begin{aligned}\mathbf{E}(\mathbf{F}|\mathbf{Y}) &= \mathbf{\Sigma}\mathbf{\Lambda}^{\mathbf{T}}(\mathbf{\Lambda}\mathbf{\Sigma}\mathbf{\Lambda}^{\mathbf{T}}+\mathbf{\Psi})^{-1}(\mathbf{Y}-\mu\mathbf{1}'-\mathbf{X}\beta) \\ &= (\mathbf{\Sigma}^{-1}+\mathbf{\Lambda}^{\mathbf{T}}\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^{\mathbf{T}}\mathbf{\Psi}^{-1}(\mathbf{Y}-\mu\mathbf{1}'-\mathbf{X}\beta) \qquad (2.14)\end{aligned}$$

Again, for the special case when $\mathbf{\Sigma} = \mathbf{I}$

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) = (\mathbf{I}+\mathbf{\Lambda}^{\mathbf{T}}\mathbf{\Psi}^{-1}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^{\mathbf{T}}\mathbf{\Psi}^{-1}(\mathbf{Y}-\mu\mathbf{1}'-\mathbf{X}\beta) \qquad (2.15)$$

2.2.2 The Maximization step

Finding the Maximum Likelihood Estimator (MLE) for the parameters β,$\mathbf{\Lambda}$,$\mathbf{\Psi}$ and $\mathbf{\Sigma}$ involves taking the first derivative with respect to each parameter and setting to 0 and solving the equation for the parameter of interest. For most of the parameter's MLE, similar forms have been encountered under various circumstances, with a little variation depending on the model set up. My model parameters are largely a mix of both the factor analysis and multivariate regression, with all parameters taking the form or a variant form of parameters from the multivariate regression. Since all parameter derivation use the same principle, I picked the parameter β and show its derivation as an example, and the rest of the parameters follow naturally. To derive the MLE for β, I combine the terms that involve β, in here the quadratic form of the likelihood function, and expand it. Note that I write everything that does not directly involve β into a compact form to avoid long notation, that is, I make the substitution $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}' - \mathbf{\Lambda}\mathbf{F}$, so that the quadratic form can be written as

$$\begin{aligned}\mathbf{\Delta} &= \mathbf{tr}((\mathbf{Y}-\mu\mathbf{1}'-\mathbf{X}\beta-\mathbf{\Lambda}\mathbf{F})(\mathbf{y}-\mu\mathbf{1}'-\mathbf{X}\beta-\mathbf{\Lambda}\mathbf{F})^{\mathbf{T}}\mathbf{\Psi}^{-1}) \\ &= \mathbf{tr}((\mathbf{H}-\mathbf{X}\beta)(\mathbf{H}-\mathbf{X}\beta)^{\mathbf{T}}\mathbf{\Psi}^{-1}) \\ &= \mathbf{tr}(\mathbf{H}\mathbf{H}^{\mathbf{T}}\mathbf{\Psi}^{-1}) - \mathbf{tr}(\mathbf{H}\beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{\Psi}^{-1}) - \mathbf{tr}(\mathbf{X}\beta\mathbf{H}^{\mathbf{T}}\mathbf{\Psi}^{-1}) + \mathbf{tr}(\mathbf{X}\beta\beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{\Psi}^{-1}) \qquad (2.16)\end{aligned}$$

Now take first derivative with respect to β

$$\begin{aligned}\frac{\partial \mathbf{\Delta}}{\partial \beta} &= \mathbf{0} - \mathbf{X^T \Psi^{-1} H} - \mathbf{X^T \Psi^{-1} H} + \mathbf{X^T \Psi^{-1} X}\beta + \mathbf{X^T \Psi^{-1} X}\beta \\ &= -\mathbf{2X^T \Psi^{-1} H} + \mathbf{2X^T \Psi^{-1} X}\beta \end{aligned} \tag{2.17}$$

where I have used the property of

$$\mathbf{tr(AB)} = \mathbf{tr(BA)} \tag{2.18}$$

$$\mathbf{trA} = \mathbf{trA^T} \tag{2.19}$$

and

$$\frac{\partial \mathbf{tr(BA^T CA)}}{\partial \mathbf{A}} = \mathbf{C^T AB^T} + \mathbf{CAB} \tag{2.20}$$

Now solve $\frac{\partial \mathbf{\Delta}}{\partial \beta} = \mathbf{0}$ to get

$$\beta = (\mathbf{X^T \Psi^{-1} X})^{-1} \mathbf{X^T \Psi^{-1} H} \tag{2.21}$$

The derivation of $\mathbf{\Lambda}$ uses the same procedure and $\mathbf{\Lambda}$ takes the form

$$\mathbf{\Lambda} = \mathbf{Y}(\mathbf{E(F|Y)})^{\mathbf{T}}(\mathbf{E(FF^T|Y)})^{-\mathbf{1}} \tag{2.22}$$

where $\mathbf{E(FF^T|Y)} = \mathbf{E(F|Y)E(F|Y)^T} + m\mathbf{Var(F|Y)}$

similarly, the MLE of $\mathbf{\Psi}$ takes the following form

$$\mathbf{\Psi} = \frac{1}{m}\mathbf{diag}((\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta)(\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta)^{\mathbf{T}} - \mathbf{\Lambda E(F|Y)}(\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta)^{\mathbf{T}}) \tag{2.23}$$

I note that the diagonal matrix $\mathbf{\Psi}$ acts as a weight of each sample, which potentially has both good and bad impacts on the parameter inference. The good thing about this is that when the samples are indeed drawn from distributions with heterogenous variance, weighting the samples is the correct approach and yields more accurate parameter estimates than otherwise. However, learning the weight of the samples is a very tricky issue and is highly sensitive to

outliers[34]. This is especially true when the parameters have to be iteratively learned with some randomly chosen initial values. When an ill conditioned vector of starting values is chosen, the result can be catastrophic, which can either lead to inaccurate solutions to the linear system or even worse, the algorithm may not converge at all.

To solve this problem, I further require $\mathbf{\Psi} = \mathbf{I}\sigma^2$. This treatment forces the ill-conditioned system back to normal by forcing all elements to have the same value, so that the effect of the outliers is minimized. To get $\mathbf{\Psi}$, I simply set each element to the average of all elements across the diagonal.

Finally, μ is set to

$$\mu = \sum_{i=1}^{m} (Y_i - X\beta_i) \tag{2.24}$$

and $\mathbf{\Sigma}$ is set to be the correlation matrix of the factor $\mathbf{E}(\mathbf{F}|\mathbf{Y})$.

2.3 Convexity of the objective function and local maxima

To show that the EM algorithm monotonically climbs the likelihood surface guarantees that it will find the global mode in my model. I first show that the objective function with the observed variable $\mathbf{F}$ is convex. Then by the general property of the EM algorithm shown by [35], we know that substitution of the unobserved variable by its expected value from its posterior distribution guarantees the algorithm to find a local maxima of the objective function, since the function has only one mode in this case, the algorithm is guaranteed to find it.

First, the likelihood function for a single sample takes on a quadratic form

$$l_1 = (y_i - \mu - X\beta_i - \Lambda f_i)^T \Psi^{-1} (y_i - \mu - X\beta_i - \Lambda f_i) \tag{2.25}$$

where Ψ^{-1} is a diagonal semi positive definite matrix, according to the following proposition

Proposition: If $F(x_1, x_2, ..., x_n) = \mathbf{x}^T C\mathbf{x}$ is a quadratic form for n variables, and if matrix C is symmetric, then F is convex $\Leftrightarrow$ C is semi positive definite

so we know that l_1 is convex. now using the following Theorem,

Theorem: If f is a function in n variables defined on a convex subset $S \subseteq R^n$, then if $f = a_1 f_1 + a_2 f_2$, where $a_1, a_2 \geq 0$ and f_1, f_2 are convex functions defined on S, then f is convex.

Clearly

$$\sum_{i=1}^{m} (y_i - \mu - X\beta_i - \Lambda f_i)^T \Psi^{-1} (y_i - \mu - X\beta_i - \Lambda f_i) \tag{2.26}$$

is convex.

Next, I show that the EM algorithm keeps climbing the hill until it reaches a local mode of the objective function (in this case the global mode), I only give some brief steps, where more details can be found in [35].

Suppose we have a likelihood function $p(\mathbf{X}|\theta)$, where $\mathbf{X}$ are the data, and θ are the parameters. The likelihood can be decomposed into

$$lnp(\mathbf{X}|\theta) = \zeta(q, \theta) + KL(q||p) \tag{2.27}$$

where

$$\zeta(q,\theta) = \sum_{Z} q(\mathbf{Z}) ln \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{q(\mathbf{Z})} \tag{2.28}$$

and

$$KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) ln \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})} \tag{2.29}$$

We see that equation 2.28 is the Kullback-Leibler divergence where $KL(q||p) \geq 0$, and the equality happens when $q(\mathbf{Z}) = P(\mathbf{Z}|\mathbf{X},\theta)$, so $\zeta(q,\theta)$ is a lower bound of $lnp(\mathbf{X}|\theta)$. In the expectation step, we set $q(\mathbf{Z})$ to be the posterior distribution, $p(\mathbf{Z}|\mathbf{X},\theta)$ to push $\zeta(q,\theta)$ as close to $lnp(\mathbf{X}|\theta)$ as possible, then the maximization step cause the log likelihood $lnp(\mathbf{X}|\theta)$ to increase, followed by the E step to push up $\zeta(q,\theta)$ to approach $lnp(\mathbf{X}|\theta)$ again..... until a local maxima of $lnp(\mathbf{X}|\theta)$, in this case the global maxima is reached.

2.4 The expected value of the factor score is a shrinkage parameter

The expected value for the random effect $\mathbf{E}(\mathbf{F}|\mathbf{Y})$ from equation 2.15 is a shrinkage parameter for a weighted ridge regression model, where we treat $\mathbf{Y} - \mu\mathbf{1}' - \mathbf{X}\beta$ as the response, $\mathbf{\Lambda}$ as the fixed covariates matrix, $\mathbf{\Psi}$ as the weight of the samples and $\mathbf{F}$ as the regression coefficients with the penalty term $||\mathbf{\Gamma F}||^2 = ||\mathbf{F}||^2$. I give a more detailed formulation of this property below.

Theorem: Consider the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{W} \tag{2.30}$$

where $\mathbf{W} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Psi})$, assume that β in the bayesian context has a prior distribution of $N(0, \mathbf{\Theta})$, then the expected value of β

$$\mathbf{E}(\beta|\mathbf{Y}) = (\mathbf{\Theta^{-1}} + \mathbf{X^T\Psi^{-1}X})^{-1}\mathbf{X^T\Psi^{-1}Y} \tag{2.31}$$

has the same form of ridge regression coefficients for $\mathbf{X}$, where the penalty term $||\mathbf{\Gamma}\beta||^2 = ||\mathbf{\Xi^T}\beta||^2$, and $\mathbf{\Xi}$ is obtained by performing a cholesky decomposition on the inverse covariance matrix $\mathbf{\Theta}$, that is

$$\mathbf{\Theta^{-1}} = \mathbf{\Xi\Xi^T} \tag{2.32}$$

Proof

First let's write the joint distribution of β and $\mathbf{Y}$ as

$$\begin{pmatrix} \beta \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} \mathbf{\Theta} & \mathbf{\Theta X^T} \\ \mathbf{X\Theta} & \mathbf{X\Theta X^T} + \mathbf{\Psi} \end{pmatrix}\right) \tag{2.33}$$

Then using the property of the joint normal distribution

$$\begin{aligned} \mathbf{E}(\mathbf{F}|\mathbf{Y}) &= \mathbf{\Theta X^T}(\mathbf{X\Theta X^T} + \mathbf{\Psi})^{-1}\mathbf{Y} \\ &= (\mathbf{\Theta^{-1}} + \mathbf{X^T\Psi^{-1}X})^{-1}\mathbf{X^T\Psi^{-1}Y} \end{aligned} \tag{2.34}$$

Where we have again used the transform equation 2.10

Now I show that the weighted ridge regression parameter estimates takes the same form. With a penalty form of $||\mathbf{\Xi}\beta||^2$, the log-likelihood of the ridge model can be written as:

$$l_c = -\mathbf{log}|\mathbf{\Psi}| - \mathbf{tr}((\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^{\mathbf{T}}\mathbf{\Psi^{-1}}) - \mathbf{tr}(\mathbf{\Xi}\beta\beta^{\mathbf{T}}\mathbf{\Xi^T}) \tag{2.35}$$

Similar to what I did to equation 2.16, we take the derivative of l_c with respect to β, which gives

$$\frac{\partial l_c}{\partial \beta} = -\mathbf{2X^T\Psi^{-1}Y} + \mathbf{2X^T\Psi^{-1}X}\beta + \mathbf{2\Xi^T\Xi}\beta \tag{2.36}$$

Set it to 0 and solve for β, to get

$$\beta = (\mathbf{X^T \Psi^{-1} X} + \mathbf{\Xi^T \Xi})^{-1} \mathbf{X^T \Psi^{-1} Y} \tag{2.37}$$

Obviously, if we set $\mathbf{\Theta^{-1}} = \mathbf{\Xi^T \Xi}$, the two takes the same form.

A special case for the theorem would be for $\beta \sim N(0, \sigma^2)$, then the expected value of β conditional on $\mathbf{Y}$ equals to the MLE of the regression coefficient under a shrinkage penalty of $\|\mathbf{\Gamma} \beta^2\| = \|I\sigma^{-1}\beta^2\|$ for a ridge model.

The fact that the $\mathbf{E}(\mathbf{F}|\mathbf{Y})$ of the random effect is a shrinkage parameter has a profound influence on how the fixed effects and the random effects should be partitioned because the non-orthogonal component can be attributed to either part. Imagine a scenario where the algorithm manages to find the correct loading matrix of the hidden factor, which is non-orthogonal to the fixed effects, however, since the estimate of the random effects is effectively shrunken toward 0, and since the fixed effects is unconstrained, this property will push the non-orthogonal components toward the fixed effects, which defeats the purpose of incorporating the random effect term to capture the hidden structure. As a result, hidden factors that contribute to false positives will not be corrected.

2.5 The model with restricted fixed effects

To address this issue, we need to either relax the constraint on $\mathbf{F}$ (or even completely cancel the constraint on $\mathbf{F}$), or give β a constraint, ideally the same constraint as $\mathbf{F}$. When assuming an unrestricted $\mathbf{F}$, $\mathbf{\Lambda}$ has to be learned in advance, then both $\mathbf{\Lambda}$ and $\mathbf{X}$ can be treated as known covariates, with the maximum like-

lihood estimator of β and $\mathbf{F}$ being obtained as in a multiple regression framework. This approach is not new as it has been suggested by [18], who used the principal components of the genotypes as the covariates. We hereby suggested a second approach, give β appropriate constraint $\|\mathbf{\Gamma}\beta\|^2$. As we have shown in equation 2.31, assigning $\beta \sim N(0,1)$ yield the same penalized regression coefficients with ridge regression. Besides, compared to the multiple regression framework where the hidden factor has to be learned in advance, this treatment allows us to seamlessly unify the multivariate regression and the factor analysis in one, which is critical for our simultaneous analysis of both the fixed effects and the hidden factors. We also get the nice property of ridge regression, e.g., 1), the shrinkage property fits in the sparse model of the eQTLs, 2), in the common scenario where $\mathbf{X}$ and $\mathbf{\Lambda}$ are usually not orthogonal especially when population structure exists, the penalized approach makes the covariate matrix $[\mathbf{X}\mathbf{\Lambda}]$ more orthogonal, leading to a more stable linear system with more accurate solutions.

Utilizing the shrinkage property of the random effect model, I propose a restricted Regression-Factor analysis model to deal with both orthogonal and non-orthogonal hidden confounding in high dimensional data analysis, which leads us to the restricted model.

2.5.1 The restricted model

The full model with the ridge can now be written as

$$\mathbf{Y} = \mu\mathbf{1}_{\mathbf{m}}^{'} + \mathbf{1}_{\mathbf{n}}\beta_{\mathbf{0}}^{'} + \mathbf{X}\beta + \Lambda\mathbf{F} + \mathbf{W} \tag{2.38}$$

with similar terms that have been described in model 2.1, except that I separate the intercept term β_0 and β here because I am restricting only β and not β_0. With

the penalty term included, the full log-likelihood can be written as

$$\begin{aligned} l_c &= -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}(\mathbf{F}\mathbf{F}^{\mathbf{T}}) - \frac{\mathbf{m}}{\mathbf{2}}\mathbf{log}|\boldsymbol{\Psi}| - ||\Xi\beta||^2 \\ &\quad -\frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}((\mathbf{H} - \mathbf{X}\beta - \boldsymbol{\Lambda}\mathbf{F})(\mathbf{H} - \mathbf{X}\beta - \boldsymbol{\Lambda}\mathbf{F})^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}) \end{aligned} \tag{2.39}$$

where $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}_{\mathbf{m}}^{'} - \mathbf{1}_{\mathbf{n}}\beta_{\mathbf{0}}^{'}$ and from the complete likelihood I can get the MLE of β as

$$\beta = (\mathbf{X}^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}\mathbf{X} + \boldsymbol{\Xi}^{\mathbf{T}}\boldsymbol{\Xi})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}(\mathbf{H} - \boldsymbol{\Lambda}\mathbf{F}) \tag{2.40}$$

and the expected values of $E(F|Y)$ as

$$\mathbf{E}(\mathbf{F}|\mathbf{Y}) = (\boldsymbol{\Lambda}^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}\boldsymbol{\Lambda} + \mathbf{I})^{-\mathbf{1}}\boldsymbol{\Lambda}^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}(\mathbf{H} - \mathbf{X}\beta) \tag{2.41}$$

Note that if I set Ξ to an identity matrix, the shrinkage on both the fixed effects and the hidden factors are the same, which seems to be a reasonable assumption when there is no strong prior information. But in the case where there is strong preference as to which one should receive more shrinkage, different $\boldsymbol{\Xi}$ can be incorporated. For example, a larger $\boldsymbol{\Xi}$ can be chosen for a more conservative estimate of the fixed effects and vice versa.

The rest of the parameters are largely similar to the unrestricted model and they are laid out below.

$$\mathbf{V}(\mathbf{F}|\mathbf{Y}) = (\mathbf{I} + \boldsymbol{\Lambda}^{\mathbf{T}}\boldsymbol{\Psi}^{-\mathbf{1}}\boldsymbol{\Lambda})^{-\mathbf{1}} \tag{2.42}$$

$$\boldsymbol{\Psi} = \frac{\mathbf{1}}{\mathbf{m}}\mathbf{diag}((\mathbf{H} - \mathbf{X}\beta)(\mathbf{H} - \mathbf{X}\beta)^{\mathbf{T}} - \boldsymbol{\Lambda}\mathbf{E}(\mathbf{F}|\mathbf{Y})(\mathbf{H} - \mathbf{X}\beta)^{\mathbf{T}}) \tag{2.43}$$

$$\boldsymbol{\Lambda} = \mathbf{Y}(\mathbf{E}(\mathbf{F}|\mathbf{Y}))^{\mathbf{T}}(\mathbf{E}(\mathbf{F}\mathbf{F}^{\mathbf{T}}|\mathbf{Y}))^{-\mathbf{1}} \tag{2.44}$$

For the global mean μ and the intercept β_0, they can be learned by iteratively evaluating these two forms until convergence

$$\mu = \frac{1}{m}\sum_{i=1}^{m}(Y_i - 1_n\beta_{0i}) \tag{2.45}$$

Where Y_i represent the i th column of Y, β_{0i} represent the intercept for the i th Y and

$$\beta_{0j} = \frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - \mu_i) \tag{2.46}$$

where I have assumed the fixed effects for the high dimensional data fit a sparse model with mean of 0.

2.5.2 Unifying the fixed and the random effects

Now that I know the penalized fixed effects take on the same form with the expected value of the random effects, I also assume that the fixed model fits a sparse model with mean of 0. These two assumptions allow me to unify the ridge regression model and the factor model into a combined framework by making the following substitutions. Let $\mathbf{\Omega} = [\mathbf{G}\mathbf{\Lambda}]$ and $\mathbf{\Gamma} = \begin{bmatrix} \beta \\ \mathbf{F} \end{bmatrix}$, I can simplify the model as

$$\mathbf{Y} = \mu\mathbf{1}_\mathbf{m}' + \mathbf{1}_\mathbf{n}\beta_\mathbf{0}' + \mathbf{\Omega}\mathbf{\Gamma} + \mathbf{W} \tag{2.47}$$

I notice that this treatment transforms the model into a simple factor analysis model, which has the same form of maximum likelihood estimator for $\mathbf{\Omega}$ as for $\mathbf{\Lambda}$, except that the first few columns of $\mathbf{\Omega}$ are fixed covariates, and the rest of the columns are for the hidden factors. Then we can use the same type of EM algorithm for the factor analysis model for inference of the parameters, with the iteration steps listed below.

1. Initialize parameter values
2. Learn μ and β_0 iteratively by using equation 2.45 and 2.46

3. Set $\mathbf{H} = \mathbf{Y} - \mu\mathbf{1}'_{\mathbf{m}} - \mathbf{1}_{\mathbf{n}}\beta'_{\mathbf{0}}$
4. $\mathbf{Var}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}} = (\mathbf{I} + \mathbf{\Omega}_{\mathbf{t}}^{\mathbf{T}}\mathbf{\Psi}_{\mathbf{t}}^{-\mathbf{1}}\mathbf{\Omega}_{\mathbf{t}})^{-\mathbf{1}}$
5. $\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}} = \mathbf{Var}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}}\mathbf{\Omega}_{\mathbf{t}}^{\mathbf{T}}\mathbf{\Psi}_{\mathbf{t}}^{-\mathbf{1}}\mathbf{H}$
6. Set the corresponding row of $\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})$ to β
7. Keep the fixed effects and the known covariates in the $\mathbf{\Omega}$ matrix fixed, update the rest using the following formula $\mathbf{\Omega}_{\mathbf{t+1}} = \mathbf{H}\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t}}^{\mathbf{T}}\mathbf{E}(\mathbf{\Gamma}\mathbf{\Gamma}^{\mathbf{T}}|\mathbf{Y})_{\mathbf{t}}^{-\mathbf{1}}$
8. $\mathbf{S}_{\mathbf{t+1}} = \frac{\mathbf{1}}{\mathbf{N}}(\mathbf{H}\mathbf{H}^{\mathbf{T}} - \mathbf{\Omega}_{\mathbf{t+1}}\mathbf{E}(\mathbf{\Gamma}|\mathbf{Y})_{\mathbf{t+1}}\mathbf{H}^{\mathbf{T}})$
9. $\mathbf{\Psi}_{\mathbf{ii}}^{\mathbf{(t+1)}} = \frac{\mathbf{tr(S)}}{\mathbf{n}}$
10. Iterate until convergence

The convergence of the algorithm can be diagnosed by checking whether the update of the likelihood or parameters approach a specified tolerance threshold. I prefer checking the tolerance of the likelihood, which can be calculated as in equation 2.39

2.5.3 The test statistics

I can use either a Likelihood Ratio Test (LRT) or Wald test to calculate the p values of the fixed effects. The LRT is performed by calculating the following value,

$$LR = -2 * (l_{null} - l_{full}) \tag{2.48}$$

where l_{null} and l_{full} correspond respectively to the log likelihood of the null model and the full model. To calculate these two log likelihood, I first take the complete model in equation 2.39, calculate the full likelihood for all $\mathbf{Y}$s given a single covariate, then I delete each Y_i at a time, for each deletion, calculate a

null likelihood. m null likelihood will be generated for all Ys for this specific covariate, which yield m likelihood ratios. Since this ratio is in expected to be asymptotically χ^2 with degree of freedom of 1, m p values can be obtained for a fixed effect.

I notice that this is a large number of EM runs for the LRT test ($m \times k$, where k is the total number of covariates), especially when very small p values need to be obtained for the ranking of the tests, in which case the tolerance of the EM have to be set higher and take even longer. So I prefer a one step Wald test.

The Wald test is performed by constructing a t test statistic like the following

$$\frac{\hat{\beta} - 0}{\sqrt{Var(\hat{\beta})}} \tag{2.49}$$

Note that I am testing the significance of only one X_i Y_j pair at a time, where $Var(\hat{\beta})$ is the corresponding diagonal element of the following form, that is, the covariance matrix of β

$$(\mathbf{I} + \mathbf{X^T \Psi^{-1} X})^{-1} \mathbf{X^T \Psi^{-1} X} (\mathbf{I} + \mathbf{X^T \Psi^{-1} X})^{-1} \tag{2.50}$$

2.5.4 Test statistics adjustment

The factor model assumes all responses Y_is are affected by the same number of factors that are preselected. While this assumption can work on data with hidden factors that affect the Ys homogeneously, it may work poorly for the real data that can be very complicated, that is, for a chosen factor number, it may fit a subset of Y_is perfectly, but for other subset of Y_is, it can either under fit or over fit, as a result, the test statistics for different Y_is may either be inflated

or deflated. To address this issue, I calculated a variance term for each Y_i, and adjusted the t statistics by the following value

$$\frac{1 + \sum_{j=1}^{n} g_i^2/\sigma}{1 + \sum_{j=1}^{n} g_j^2/\psi} \tag{2.51}$$

where g is a fixed covariate of interest, σ is the error variance for a particular Y_i and ψ is the global variance, where we have assumed $\Psi = I\psi$. This effectively takes away the skewness of the p values by taking the over fitting or under fitting of the model into account, as a result, the p values that are either inflated or deflated will be bring back to uniform.

2.5.5 Selecting the factor numbers

Various techniques can be used to select the factor number for the model. For example, the Akaike information criterion criteria which is calculated as,

$$AIC = 2k - 2ln(L) \tag{2.52}$$

or the Bayesian information criterion (BIC), which is calculated below

$$BIC = -2ln(L) + Kln(n) \tag{2.53}$$

where K is the parameter number, n is the sample size and L is the likelihood. Although I found these two criteria works well for a relatively simple data structure, it shows less satisfactory performance for more complicated ones that have multiple layers of hidden structures. Thus I prefer selecting the factor number by manually examining the eigen spectrum of the data. The eigen spectrum is obtained by first do a Principal Component Analysis on the data, then the variance proportion that is explained by each component is calculated. The appropriate number of factor number was selected based on how well each eigen vector can be visually separated from the rest.

2.5.6 Summary

To control for the false positives and increase the power to detect true fixed effects associations in a high dimensional data set, I propos a statistical model that combines the multivariate regression and the factor analysis models to control for the hidden factors. I discovered the dynamics of the fixed and random effects by noticing that the expected value of the hidden factor is a shrinkage parameter. Based on this observation, I proposed and constructed a shrinkage version of the model that unifies the multivariate ridge regression and factor analysis. In the next chapter, I show the performance of the model by extensive simulations and applying it to an interesting data example.

CHAPTER 3

HEFT: EQTL ANALYSIS OF MANY THOUSANDS OF EXPRESSED GENES WHILE SIMULTANEOUSLY CONTROLLING FOR HIDDEN FACTORS

3.1 Introduction

Studies that have identified expression Quantitative Trait Loci (eQTL), the genetic loci that produce variation in cellular or tissue gene expression levels, have demonstrated that a considerable fraction of gene expression variation has a genetic basis [36, 7, 37, 38, 39, 40, 41]. Recently, more precise measurement of genome-wide gene expression levels using RNA-Seq technology [6], combined with greater marker coverage of genomes, has increased the resolution of eQTL analyses and has allowed more precise dissection of eQTL effects[42]. What's more, a spectrum of new genome-wide assays making use of next-generation sequencing, such as Methly-Seq [43] and CHiP-Seq [44], are providing quantitative data on other cellular profile variables that can be analyzed using an eQTL approach, opening the door for a broader xQTL framework [45, 46]. This expanded capability and diversity of eQTL detection has also been accompanied by an appreciation that eQTL can provide useful insights into the genetic basis of disease [47]. For example eQTL identification is now routinely incorporated into the analysis of disease risk and other complex aspects of physiology [36, 7, 47, 48, 49]. A consequence of these trends is a renewed interest in analysis methodologies used to identify eQTL from genome-wide data [32, 24, 26, 25, 27, 28, 29, 30, 31]. For these new methods, there is a premium on the ability to identify as many eQTL as possible while simultaneously providing strict false positive control. High performing, fast, and reliable methods

will also be particularly valuable for analyzing the highly multivariate, mixed data-type xQTL studies that are on the near horizon.

For a typical eQTL study that includes genome-wide data on both gene expression and genetic markers, identification of eQTL is typically accomplished using standard linear modeling approaches, where marker genotypes with a significant association with one or more expression variables are assumed to either indicate an eQTL or a marker that is in linkage disequilibrium with the eQTL polymorphism, i.e. the marker indicates the local genomic position of an eQTL [50, 51]. While such approaches are straightforward and successful, it is well appreciated that factors responsible for variation in gene expression, if unaccounted for in the statistical model, can dramatically affect both power and precision of genome-wide eQTL detection [51]. This is particularly true in uncontrolled study designs, as is often the case with human eQTL studies, where unmeasured environmental and other factors can influence gene expression profiles and confound eQTL analysis [52, 53]. More precisely, if the effects of unaccounted for, factors on gene expression are orthogonal to effects of eQTL, the factors contribute to the error term and this reduces power to detect eQTL. If the effects of unaccounted for factors are non-orthogonal to the effects of eQTL, the result can be a false positive [18].

That unaccounted for factors can be a problem for eQTL identification is not surprising given the many studies demonstrating that gene expression levels are highly variable and depend on a host of genetic [40, 41] and non-genetic factors [52, 53]. For statistical modeling purposes, we can categorize expression factors into three cases that require different analysis approaches: (1) a factor

that is well represented by a variable that is directly measured in the study, (2) a factor that can be inferred from the genotype data, and (3) a factor with effects that can be learned from gene expression data. The first includes cases where the measured variable, such as experimental batch, an environmental indicator, gender, a disease state of an individual, etc. can be directly incorporated into the statistical model as a covariate. The second includes factors such as cryptic population structure [18] or relatedness among individuals that can produce variation in measured gene expression levels. For many of these cases, appropriate variables can be inferred directly from the genome-wide genotype data, which can then be secondarily incorporated as fixed or random covariates to correct for factor effects [18, 54]. The third case includes expression factors that cannot be well modeled with covariates inferred from genotype data but have effects that can be learned from the covariance among expressed genes [24, 25]. For this case, the assumption is that the expression factor effects are large enough that the effects of the factors, although not the factors themselves, can be learned using a factor analysis or related approach [26, 25]. These learned factors can then be incorporated into the eQTL analysis as covariates [26] and the eQTL analysis can be conducted on the residuals of the expression variables after subtracting the learned factors [25]. The value of accounting for factors in an eQTL analysis that can be learned from expression covariance is just beginning to be appreciated and several recent methods have been proposed for this purpose [32, 24, 26, 25, 27, 28, 29, 30, 31]. We note that these publications variously refer to these expression factors as hidden confounders [24], non-genetic factor [25], surrogate variables [27] etc., but here we refer to them as hidden factors.

In this chapter, I construct a new method for eQTL analysis that accounts for

the effects of hidden factors: Hidden Expression Factor analysis (HEFT) based on the regression-factor model proposed in the last chapter. The HEFT framework unifies a number of desirable goals when performing an eQTL analysis in presence of hidden factors: p value identification of eQTL with individual or pleiotropic (multiple) effects on expressed genes when using a ridge penalty, the ability to learn both orthogonal and non-orthogonal hidden factors that can inflate or deflate p values, and efficient scaling, such that an eQTL analysis of thousands of gene expression variables and hundreds of thousands of marker genotypes can be completed in a few hours on a standard desktop. Quite critically, HEFT simultaneously assesses eQTL while learning hidden factors without the need for any type of pre-estimate of factor effects, the importance of which I illustrate by comparing the performance of HEFT to two-step and related hidden factor methods when analyzing simulated data [25, 24, 31]. I also demonstrate the real-world discovery value of a hidden factor analysis by using HEFT to identify eQTL that affect gene expression in the human lung of a sample of smokers and nonsmokers by assessing possible associations of 7,575 expression variables with 191,959 genotypes. In this analysis, I both empirically validate the hidden factor detection of HEFT, by recovering the effects of smoking when this covariate is treated as missing, and I use the full HEFT model to identify eQTL that could not be found by standard eQTL analyses. Many of these newly discovered eQTL have clear connections to lung physiology and disease, including a *cis*-eQTL for MTRR *cis*-eQTL for GTF2H1, two genes that have been independently associated with lung cancer [55, 56], a *cis*-eQTL for RUVBL1, a gene that is over-expressed in the tumor cells of several tissue including the lung[57], a *cis*-eQTL for TEFM, also known as C17orf42, where inactivation of TEFM leads to respiratory incompetence[58], and several eQTL with

pleiotropic *trans*-effects on genes associated with lung phenotypes.

3.2 Methods

3.2.1 The HEFT Model

The HEFT framework assesses associations between genotypes and expression variables by employing the same model as shown in chapter 2, equation 2.38. In this framework, I assume that expression variables have been scaled to a common variance, I also assume complete expression and genotype data or that missing values have been imputed prior to analysis. When considering additional fixed covariates, whether directly measured or independently inferred from genotypes (e.g. populations structure), these are incorporated 2.39 as additional fixed effects. In addition, additive, dominance, and the simultaneous effects of multiple genotypes (including epistasis) can be handled in this framework, although I restrict the current treatment to assessing a single genotype at a time using an additive coding.

A number of proposed methods make use of the modeling strategy of equation 2.39 by applying a two-step approach, where hidden factors are learned from a separate factor analysis and the inferred loadings are then incorporated into a fixed $\mathbf{\Lambda}$ [26] to adjust the p values. In my treatment here, I simultaneously infer genotype associations and learn factors (i.e. I use an unrestricted $\mathbf{\Lambda}$) by imposing constraints on $\mathbf{F}$ and β to account for lack of identifiability of the combined genotype and hidden factor effects. We do this by introducing a hierarchical

control by assuming $\beta \sim N(0, \sigma_\beta^2)$ and $\mathbf{F} \sim N(0, \sigma_F^2)$. With this approach, the form of expected value of the parameters from its posterior distribution has the same form as regression coefficients obtained by a ridge regression with the penalty term $\|\mathbf{\Gamma}\beta\|^2$ equal to $\|\mathbf{I}\sigma_\beta^{-1}\beta\|^2$. This hierarchical approach therefore places a ridge penalty on both the genotype and factor, which will be appropriate when we expect the genotype and factor effects to follow a relatively sparse model when considering the entire variable set, a reasonable assumption in many cases when the expression variable set m is large. This approach also has the additional benefits of a ridge regression e.g. stable solution on the non-orthogonal linear equations, smaller variance of the estimator of βs, asymptotically correct estimates of parameters, etc.

We note that since $\mathbf{\Lambda}$ is unrestricted, setting the value of σ_F^2 to a constant has no effect on the results so we set $\sigma_F^2 = 1$ in practice. We also find that setting σ_β^2 to a value that is the same or higher than the scaled variance of the expression variables (i.e. such that the hierarchical control is diffuse), there is no qualitative effect on results. However, we do need to shrink β and $\mathbf{F}$ by the same amount to address the identifiability issue caused by the non-orthogonal factors. We therefore adopt this approach in our analyses by scaling the variances of expression variables to 1 and set $\sigma_\beta^2 = 1$, such that $\sigma_F^2 = \sigma_\beta^2$, an approach that prevents biasing estimates towards genotype or factor effects when these are non-orthogonal and also has convenient properties for reducing the computational complexity of the EM algorithm (see next section and supplement). We also note that the lack of a unique solution for $\mathbf{\Lambda F}$ is not an issue for our treatment, since we are only interested in accounting for the overall effects of hidden factors and not in learning either factor loadings or the factor scores.

3.2.2 Likelihood and EM algorithm

For the HEFT model, the complete likelihood has the same form with equation 2.39 from the previous chapter. For the purposes of eQTL analysis, we are only interested in the estimates of the β for a given marker, where we use the parameter estimate of β which take the same form with equation 2.40

Obtaining these estimates is accomplished using an Expectation-Maximization algorithm, which has time complexity scaling $O(nmp^3)$, where n is the sample size, m is the number of genes and p is number of the factors (see appendix and supplementary materials). As p is small, the algorithm is extremely efficient. What's more, the likelihood function of the complete model is convex (see previous chapter) and since all expression variables are analyzed simultaneously, analysis of an individual genotype and all expressed genes can be done in a single step with a single run of the algorithm.

3.2.3 Selection of factor number

We note that the true number of hidden factors in the model p can never be known with certainty. While for simulated data, standard model selection approaches such as Akaike information criteria (AIC) or Bayesian information criterion (BIC) can be used to correctly infer the number of factors (see previous chapter), we have found that for real data, these can select too many factors, resulting in clear hallmarks of data over-fitting. In practice, we therefore select the number of factors by assessing the eigenspectrum of the overall gene expression covariance and we include up to k factors that are visually distinguishable from

the rest of the eigenvectors, and in cases where known covariates are included, this factor number should be reduced correspondingly. This approach performs well for simulated data and for real data, producing a reasonable enrichment of significant eQTLs without over-inflating the genome-wide distributions of p values as measured by the genome-wide inflation factor λ [59] (see below). Thus, while the $p < n$ selected factors will not necessarily provide a perfect fit of covariance, selecting the number of factors to provide a reasonable fit to observed covariation provides a good approximation in practice.

3.2.4 P values and identification of eQTL

As with a standard eQTL analysis, identification of eQTL using HEFT is accomplished using p values. The HEFT software can calculate several test statistics, including an asymptotically exact likelihood ratio test (LRT) (see chapter 1). Given that the calculation of the for both the null and full model for the LRT requires m runs of the EM algorithm per marker to assess each genotype-gene pair, for the purposes of speed, we favor the t test statistic, which requires one run of the algorithm per marker (see chapter 1). While this test is not asymptotically exact, we find this to perform well in practice, where resulting p values are uniform under the null and the statistic has comparatively good power (see below).

The test statistics can be constructed for each individual genotype-expression pair, for subsets of the m expression traits, or for the entire set of m traits, so they can be used to provide an overall assessment of whether the genotype indicates an eQTL by assessing all m expression variables or any subset. However,

since a single significant genotype-gene association indicates an eQTL, in this treatment, we follow the standard practice of eQTL analysis and assess one pair relationship at a time and interpret rejection of the null for at least one pair as evidence of an eQTL. The pleiotropic effects of an eQTL (i.e. the effects on multiple genes) are determined by the set of genotype-gene pairs for which the null is rejected for a genome-wide multiple test corrected significance threshold.

3.2.5 Connections between HEFT and other eQTL hidden factor methods

A number of proposed methods use a two-step approach for hidden factor eQTL analysis, where hidden factors are learned from a separate factor analysis and either the inferred loadings Λ are incorporated as covariates [26] or the residuals $Y - \Lambda F$, which are assumed to be free of the hidden factor structure, are used to perform secondary eQTL analysis [25]. Heuristic simultaneous hidden factor methods have also been proposed, such as Surrogate Variable Analysis (SVA) [27], which uses a re-weighted surrogate variable analysis to partition the response matrix into genotype and factor related subsets, where these partitions are iteratively updated. In our treatment here, we avoid the inherent problems of two-step and heuristic procedures by simultaneously inferring eQTL and hidden factors using a likelihood approach.

HEFT is most similar to the simultaneous linear mixed model (LMM) approach of Listgarden et al. [24, 31]and the variational Bayesian (VBQTL) approach of Stegle et al. [25]. Unlike the LMM approach to hidden factor learning, HEFT

is convex and does not require to use the conditional likelihood to integrate the hidden factor out, we can produce an equivalent model formulation to the LMM model using the HEFT framework by not applying a ridge penalty to the genotype and by including a ridge penalized population structure factor and use a conditional likelihood. Unlike the VBQTL approach, HEFT uses a frequentist approach instead of approximate Bayesian objective function and there are some differences in how the underlying model is specified (e.g. while VBQTL does use a ridge penalty, additional conjugate priors on the variance terms are employed that we do not include). We also note that while the VBQTL framework can in theory apply a simultaneous eQTL / hidden factor analysis, the released R package does not support simultaneous analysis.

3.3 Simulations and Data

3.3.1 Simulated Data and Analyses

I simulated data for each of the following scenarios: a) no eQTLs and no hidden factors (null scenario 1), b) no eQTLs with hidden factors (null scenario 2), c) eQTL where each affects one expressed gene (no pleiotropy) and no hidden factors, d) a combination of pleiotropic and non-pleiotropic eQTL and no hidden factors, e) non-pleiotropic eQTLs with hidden factors, f) a combination of pleiotropic and non-pleiotropic eQTL with hidden factors. For each of the scenarios with hidden factors (b, e, f), I simulated 10 datasets where the hidden factors effects were orthogonal to the entire set of markers and 10 datasets with hidden factors that were non-orthogonal to a non-trivial subset of the markers.

For the scenarios with no hidden factors (a, c, d), I also simulated 10 datasets each. The sample size for each dataset was fixed at n=200.

To generate the genetic markers of each dataset, SNP genotypes were generated using the coalescent simulator MaCS [60] using the default approximation for tree width. We simulated 5 Mb of marker data for a single diploid populations of size $N_e = 10,000$, with a population mutation rate of $\theta = 4N\mu = 0.001$ and the recombination rate of $\rho = 4N\kappa = 0.00045$, values taken from Voight et al. [61]. For a dataset, we randomly selected 1000 SNPs from those with a derived Minor Allele Frequency (MAF) greater than 0.1, producing average linkage disequilibrium of 0.45 ± 0.01 for all ten datasets for pairwise markers measured by r^2. We note that we did not include population structure in our simulation analyses as we were interested in assessing the ability of hidden factor methods to detect eQTL without this additional layer of complexity. Again, we note that HEFT can include a fixed effect correction for population structure and we use this approach for the analysis of real data (see below).

To generate the gene expression values of each dataset, I simulated 500 gene expression variables with standard normal error. For the eQTL scenarios with no pleiotropy (c and e) we randomly selected 50 uncorrelated markers to be eQTL, where the additive effect of each on a randomly selected gene was drawn from a standard normal. For the cases with pleiotropy (d and f) we included 50 eQTL with individual gene effects and selected an additional 20 uncorrelated SNPs each influencing 20 expression variables each, where again, the effect on each gene was selected from a standard normal. Overall, the total variation explained by the eQTLs for a given gene ranged from 5.5e-07 to 0.92, with the vast majority

in the range of 0-0.025 (see supplement). For each dataset with hidden factors (b, e, f), we additionally incorporated the effects of four factors. To simulate a non-orthogonal factor, the scores of individuals on the first principal component of the correlation matrix of 100 randomly selected markers was used to assign individuals into five total groups, where the individuals with the largest 40 scores were assigned to group 1, individuals with the next largest 40 were assigned to group 2, etc. (i.e. factor effects were orthogonal to each other although non-orthogonal to the 100 SNPs). For each group, a single effect was then assigned drawing from $N(0, 1)$ or from $N(0, 3)$. For orthogonal factors, we applied the same procedure but randomly assigned each individual to one of the five groups. While for the latter, this does not prevent a factor being non-orthogonal to some markers, we found that each factor was approximately orthogonal to almost all of the markers in the dataset in practice.

We analyzed each simulated dataset with the following eQTL methods: a naive linear regression method, a two-step hidden factor analysis method, the mixed model approach LMM [24], the variational Bayes method VBQTL [25], and HEFT. With each method, we analyzed one SNP at a time for eQTL associations. For the two-step analysis, we estimated factor structure using a factor analysis of the expressed genes and then used the residuals $Y - \Lambda F$ to do a secondary analysis, i.e. we applied a two-step approach within the HEFT framework (referred to as HEFT-TS). For LMM [24], we could not get the software provided by the authors to work, so we re-implemented the algorithm. We note that we used the same convergence and other implementation criteria as described by the authors [24] and that our implementation performed as they described. We also note that we did not make use of their populations structure component to

allow for an appropriate comparison because there is no population structure in our simulated data (i.e. we applied LMM-EH and not LMM-PS-EH). While VBQTL [25] can in theory perform simultaneous analysis of eQTL and hidden factors, there are no simultaneous inference components implemented in the available R software package and requests to the authors for software capable of simultaneous analysis were unsuccessful. We therefore applied VBQTL using their two-step option. For all analysis methods where the factor number could be controlled (HEFT-TS, VBQTL, and HEFT), we analyzed each scenario where there were no hidden factors (a, c, d), we analyzed each dataset with factor number p=1, 2, and for each scenario where there were four hidden factors (b, e, f), we analyzed each dataset with factor number 4, 5, 7.

For each analysis method, the association of each SNP-gene expression pair was assessed. For the naive regression and HEFT we used the resulting *p* values. For HEFT-TS and VBQTL, we followed the same procedure as applied in the VBQTL paper [25], where we extracted a *p* value like statistic from a linear regression model applied to the residuals after fitting the factor model. For LMM, we calculated the *p* value statistic as described in their paper [24]. For assessing performance, a *p* value below a selected threshold for a SNP-gene pair representing a true eQTL was counted as a true positive and similarly, *p* values below the selected threshold for a SNP-gene pair that was not an eQTL were counted as a false positive (and similarly for true negatives and false negatives). We note that while linkage disequilibrium was not overly strong in our simulated marker datasets, with this approach, non-eQTL SNPs that were in strong linkage disequilibrium with eQTL SNPs could contribute multiple false positive signals. Thus, while we potentially counted a few cases as false positives

that would be merged into a single 'true' positive in a real eQTL analysis (where the true eQTL are not known), by applying the same conservative criteria for all analysis methods, this provided a common and fair comparison of performance.

3.3.2 Lung Airway Dataset

We used HEFT to identify eQTL affecting gene expression in the lung Small Airway Epithelium (SAE) using a dataset that included 79 smokers and 37 non-smokers recruited from the New York City area. The individuals in the sample were of different genders, different ancestry groups, and were characterized as non-smokers or smokers and were further labeled as healthy or having a lung disease phenotype (see Table 3.1, and the supplementary material for complete details of the demographic information). Details concerning data collection for these samples have been provided elsewhere [52]. Briefly, SAE cell populations were collected by bronchial brushing of the small airway [62] and RNA was hybridized to the HG-U133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA) using standard protocols. To avoid the problem of probe sets mapping to wrong genes, we used the custom mapping provided by [63] and the Robust Multi-array Average (RMA) [64, 65] normalization method to convert array probe expression measurements into a single expression measurements for genes with unique Ensemble numbers. We further removed genes with individual expression values beyond 3 standard deviations of the mean, which appeared likely to be outliers. This provided data on ~7575 protein-coding genes, an unknown subset of which are operating in the regulation and response behaviors of the pulmonary environment.

Blood was also collected from each individual and Affymetrix 500k microarrays were used to provide SNP genotypes. After filtering SNPs with a MAF below 0.1, significant deviations from Hardy-Weinberg equilibrium as assessed by a p value < 0.05 for an efficient exact test [66], and those genotypes with any missing observations using PLINK [67], this left 191,959 genotypes for analysis. The complete expression and genotype dataset analyzed in this study have been deposited in NCBI's Gene Expression Omnibus [68] and are accessible through GEO Series accession number GSE32030 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=xtyrvgaswmoquzo&acc=GSE32030`).

We initially applied the factor component of the HEFT model to just the expression data treating the known smoking covariate as missing information to assess the recovery ability of the factor component of the model. We then applied a complete HEFT analysis to the entire dataset. For this analysis, we selected the hidden factor number by visually examining the eigen spectrum of the gene expression correlation matrix and selected 5 factors that are clearly separable from the rest. To account for the obvious population structure in these data, we applied a factor analysis to the genotype covariance matrix [69] and incorporated the loadings of the first factors as fixed covariates, where this number was selected from the genotype covariance matrix eigen spectrum. We additionally included fixed covariates including gender, disease status and the smoking status. For binary covariates such as gender and smoking, we encoded them as 0 and 1, while three level disease status was encoded as a $n \times 2$ binary design matrix of either 0 or 1. For a baseline comparison, we also applied a multiple regression model including all of the same fixed covariates. A Bonferroni corrected threshold 0.05 / (7575 $\times$ 191959) = 3.438578e-11 was used to assess

significance of each SNP-expression pair.

Table 3.1: Population demographics of the human lung airway epithelium study.

	Small airway epithelium[1]		
	Healthy nonsmokers[2]	Healthy smokers[3]	Smokers with pulmonary disease[4]
# of Samples	38	57	21
Gender (M/F)	28/10	35/22	18/3
Ethnicity (B/W)[5]	22/16	43/14	8/13
Age[6]	43±12	45±9	55±9
Pack-year history[7]	0	28±17	45±27

1. Data are presented as mean ± standard deviation where appropriate.
2. RNA samples collected by bronchial brushing of the 10-12th bronchial order[52].
3. Life-long nonsmokers with normal lung functions as measured by spirometry and diffusion capacity of carbon monoxide[52].
4. Current or ex-smokers with normal lung functions as measured by normal spirometry and diffusion capacity of carbon monoxide[52].
5. Current or ex-smokers with pulmonary disease as defined by their lung functions: either with Chronic Obstructive Pulmonary Disease (COPD) as defined by the GOLD criteria [52] or early emphysema as defined by normal spirometry and reduced diffusion capacity of carbon monoxide (<80%) [52].
6. African American (B=Black) or Caucasian (W=white).
7. Presented as mean ± standard deviation.
8. Calculated for each individual as the number of packs of cigarettes smoked per day times the number of years of self-reported smoking history presented as mean ± standard deviation.

3.4 Result

3.4.1 Performance for null and standard eQTL scenarios

For datasets simulated under scenario a, where there are no eQTL and no hidden factors (null scenario 1), all five eQTL analysis methods returned a uniform distribution of p values for the set of all SNP-gene tests as measured by genomic inflation factor in a range of 0.99-1.01 [70, 71], indicating they all performed appropriately for this null scenario (see supplemental figures and table). This outcome was observed regardless of the number of factors that were provided to HEFT-TS, VBQTL, and HEFT, indicating that these methods are also robust to incorporating the wrong number of factors ($p = 1, 2$) for this null scenario.

For datasets simulated under scenario b, where there are no eQTL and with hidden factors (null scenario 2), we considered performance for cases where the effects of the four hidden factors were (approximately) orthogonal to all SNPs and cases where the effects of the four hidden factors were non-orthogonal to 10% of the SNPs. For the orthogonal case, with the correct factor numbers all methods including SLR produced an almost uniform distribution of p values (λ=1-1.02). This result seemed to be robust for all methods but HEFT, where the later shows inflated p values when a too small factor number is offered ($p = 3$). This is not surprising since the HEFT model tried to find a hidden factor configuration that maximize the genetic effects, which can only find a false one when the chosen factor number is too small, which as a result, falsely increased the magnitude of the genetic effects. This false structure will not affect other methods because they don't try to find the best configuration of the genotypic ef-

fects and the hidden factors, for example, for the two step approach, the hidden factors were inferred separately so that the false structure can subtract equal amount of genotypic effects, and for LMM-EH, the hidden factors were integrated out so that the false structure contribute to the error term and the p value distribution remain unchanged. For factor number of ≥ 4, all methods performs well.(see supplemental figures and tables).

For the case of non-orthogonal hidden factors under this same null scenario, the performance issues for naive regression diverged far from the null expectation where far too many small p values were returned, a result that in practice would result in a large number of false positives (Figure 3.1 and see supplemental table). This result is expected given that the naive linear regression is unable to distinguish an eQTL signal from the effects of hidden factors. All other methods except VBQTL returned p values conforming to the null expectation for the non-orthogonal case when provided the correct or greater than the true number of factors ($p \geq 4$), again indicating that these methods perform appropriately in this null scenario. VBQTL require a factor number of $p > 4$ to perform well. HEFT-TS and HEFT also lost their robustness for factor number of $p \leq 3$ because they can't find the correct structure any more. We also see that as the number of factors increase, the two step approach shows slightly deflated p values due to the overcorrection, and HEFT shows slightly inflated p values because of its property of retaining the genotypic effects without overcorrection.

For the standard eQTL scenarios where there are eQTL with individual gene effects (no pleiotropy, scenario c) or with a combination of individual and pleiotropic effects (pleiotropy, scenario d) but where there are no hidden fac-

tors, the naive linear regression and HEFT had the best performance (Figure 3.2 and see supplemental figures), where this was the case regardless of factor number provided to HEFT ($p = 1, 2$). HEFT-TS and VBQTL had slightly worse performance than regression and HEFT in the case of no pleiotropy and noticeably worse performance in the case of pleiotropy.

Given that the linear regression statistical model was appropriate in the standard eQTL scenario where there are no hidden factors, this good performance was expected. HEFT was also able to correctly control the hidden factor effects to negligible values, producing performance on par with linear regression. Since HEFT-TS and VBQTL estimated hidden factor effects in a separate step from the eQTL analysis, they tended to account for variation due to eQTL with the hidden factors (i.e. these methods over-fit the hidden factor effects) and therefore had lower power, where this outcome was more pronounced as the provided factor number was increased. This over-fitting result was also more pronounced in the scenario where there are pleiotropic eQTL, which therefore resemble the effects of hidden factors. We note that the performance of LMM was far worse than any of the other methods (for all scenarios), likely because the high-dimensional hidden factor control of the mixed model was resulting in even more extreme over-fitting of variation that was due to eQTL (we provide a intuitive explanation for this over-fitting problem with LMM in the previous chapter). In sum, in the scenarios where there are no hidden factors, all of the hidden factor methods except for HEFT had significantly lower power compared to a naive linear regression.

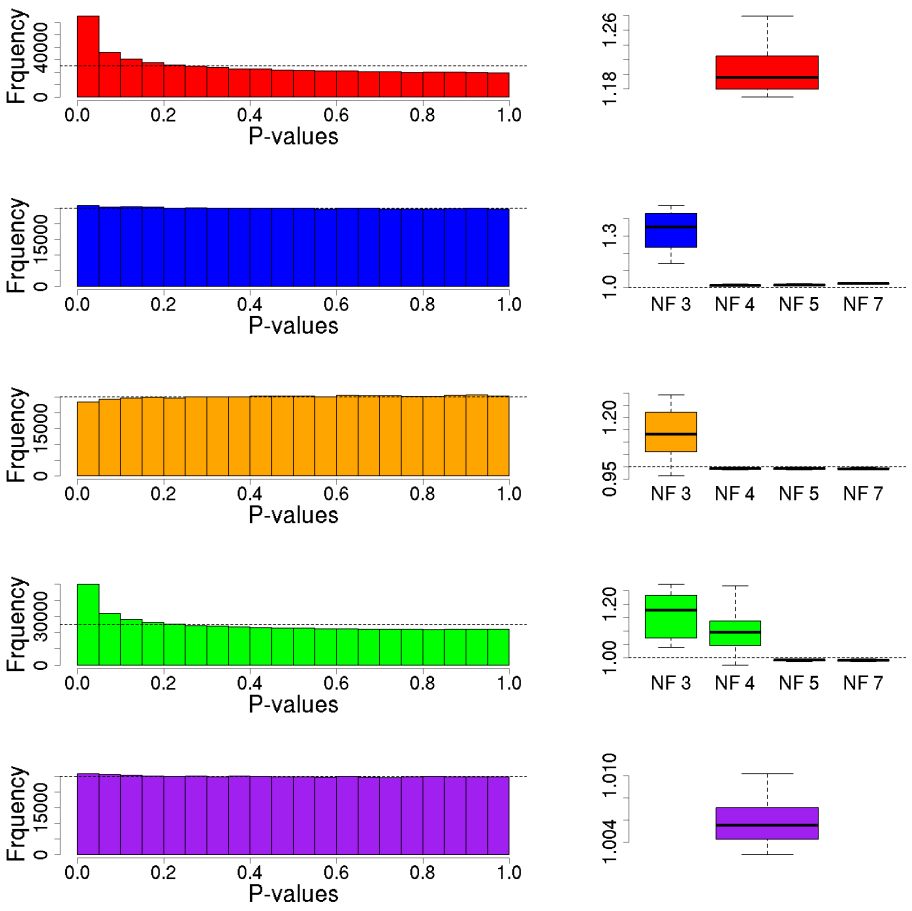

Figure 3.1: Histograms and boxplots showing the distributions of p value for all SNP-gene test of association for a non-orthogonal scenario with no eQTL and with hidden factors that are non-orthogonal to 10% of the SNPs(scenario b). The left column shows the histogram of the p values for a specific simulation with factor number of 7 (when the factor number applies), and the right column shows the boxplots of the inflation factor for p values of all ten simulations. From top to bottom are respectively linear regression, HEFT, HEFT-TS, VBQTL and LMM-EH. For the boxplots, from left to right corresponds to factor number of 3,4,5,7 when factor numbers apply.

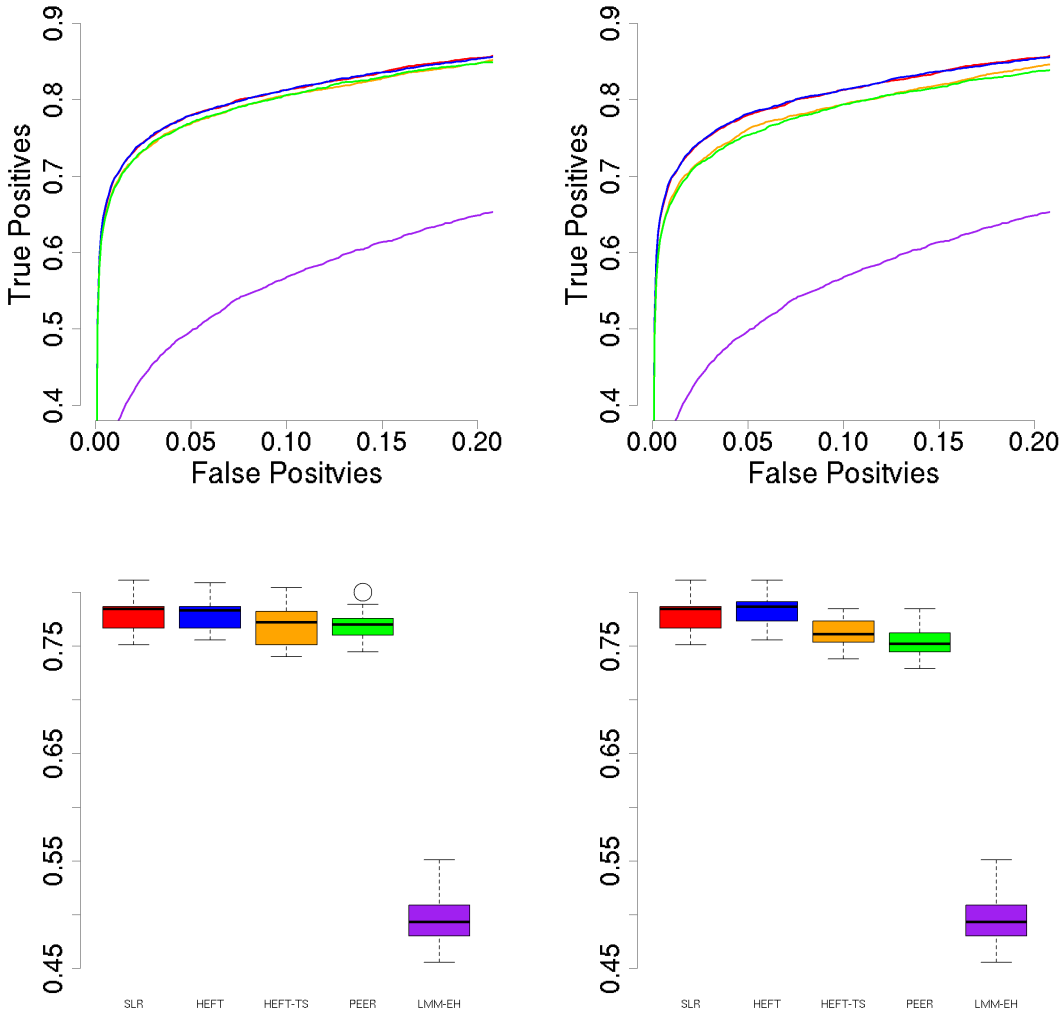

Figure 3.2: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenario where there are pleiotropic eQTLs but no hidden factors (scenario d), where the left and right columns correspond to provided factor numbers of 1 and 2. The methods are color coded the same as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

3.4.2 Performance for eQTL and hidden factors.

For the scenarios where there are both eQTL and hidden factors, performance depended heavily on whether there was no pleiotropy (scenario e) or pleiotropy (scenario f). In the scenario where none of the eQTL had pleiotropic effects, all hidden factor methods performed better than naive linear regression (Figure 3.3 and see supplemental figures). This was the case regardless of whether the hidden factor effects were orthogonal or non-orthogonal although the effects was more pronounced in the non-orthogonal cases, mostly because the non-orthogonal case has more genotypes correlated with the factors, yielding more false positives. Overall, HEFT, HEFT-TS, and VBQTL had approximately the same performance when provided the correct number of factors or more than the correct number ($p = 5, 7$), indicating that they were learning the hidden factor effects correctly (although these methods showed slightly reduced performance the larger the number of provided factors). we note that HEFT shows reduced performance when the factor number is too small than the real number($p = 3$), where again, the false configuration of the genotypic effects and the hidden factors can cause a larger number of false positives than expected and VBQTL shows reduced performance from a reduced factor number of 4. While LMM performed better than naive linear regression, the overall performance was worse than all the other hidden factors methods (regardless of the factor number provided), again likely due to the over-fitting of the underlying mixed model.

For the scenario where there are eQTL with pleiotropic effects and hidden factors, HEFT noticeably outperformed all other methods (Figure 3.4 and see supplemental figures) as long as the given factor is not too small. Again, this

was the case regardless of whether the hidden factor effects were orthogonal or non-orthogonal, although again, the effects was more pronounced in the non-orthogonal cases because of the increased number of false positives. This result is likely due to HEFT-TS, VBQTL, and LMM accounting for many of the pleiotropic effects of eQTL as hidden factors and they therefore have reduced power to detect eQTL pleiotropic effects. In fact, the reduced performance of LMM due to this effect was so extreme it no longer performed better than the naive linear regression.

3.4.3 Recovery of the smoking factor when treated as hidden.

As an empirical assessment of the ability of HEFT to recover hidden factors, we used the factor learning component of HEFT to analyze the lung SAE gene expression data, where the known information about whether individuals were smokers or nonsmokers was treated as missing. Smoking has a well-characterized effect throughout the SAE transcriptome [52] so when treated as a missing covariate, this factor should be learnable from the observed expression variation.

Figure 5 shows a plot of the samples projected on to the first two factors and the factors two and three that were recovered by HEFT, where smokers are presented in red and nonsmokers in blue. These plots provide a good visual separation of smokers and nonsmokers. Thus, the factor component of HEFT was able to learn the effects of smoking status when this covariate was treated as a hidden factor. From this analysis, it appears the influence of smoking is more complex than could be well modeled with a single fixed covariate for smoking

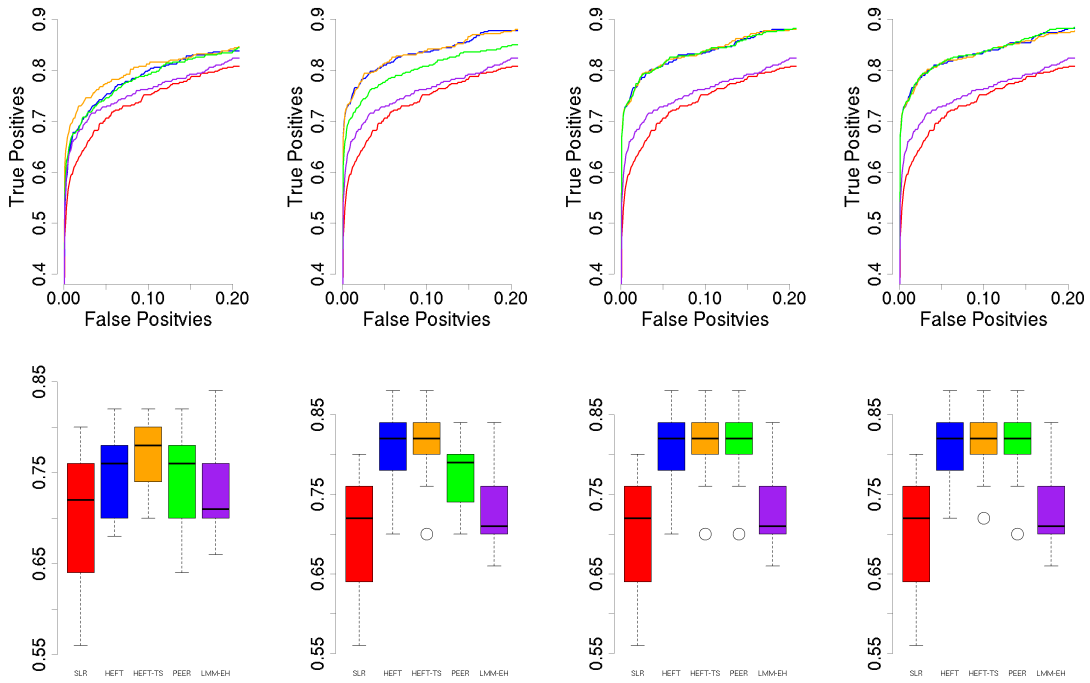

Figure 3.3: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenarios where there are eQTL effects and non-orthogonal hidden factors (scenario e), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively. The methods are color coded the same as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

status. This indicates that even in the unusual case where the critical factors are known and measured, it may be of value to learn hidden factors in an eQTL analysis.

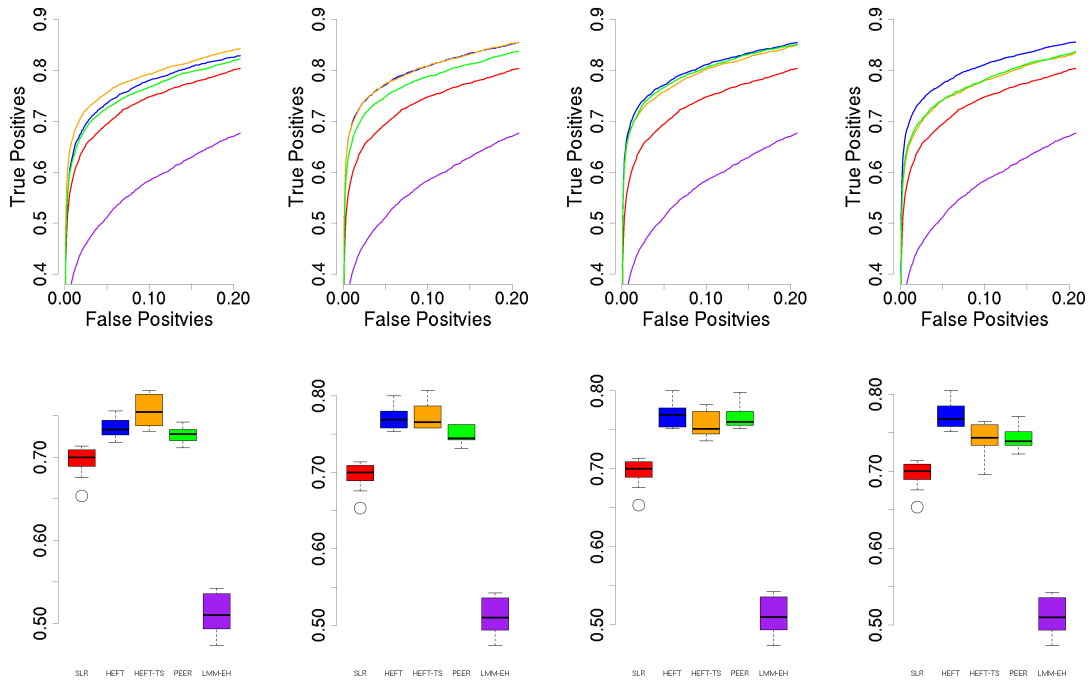

Figure 3.4: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenarios where there are pleiotropic eQTL effects and non-orthogonal hidden factors (scenario f), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively. The methods are color coded the same as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

3.4.4 Identification of lung airway eQTL using HEFT

We used both HEFT and, for comparison, a naive linear regression to analyze the 7575 SAE expressed genes and 191,959 marker genotypes using a hidden factor number of p=5 and appropriate covariates. The entire analysis took about 19 hours on a 8 core 2.6G processor. After ranking the full 7575 $\times$ 191959=1.45e9 p values for both HEFT and naive linear regression, we

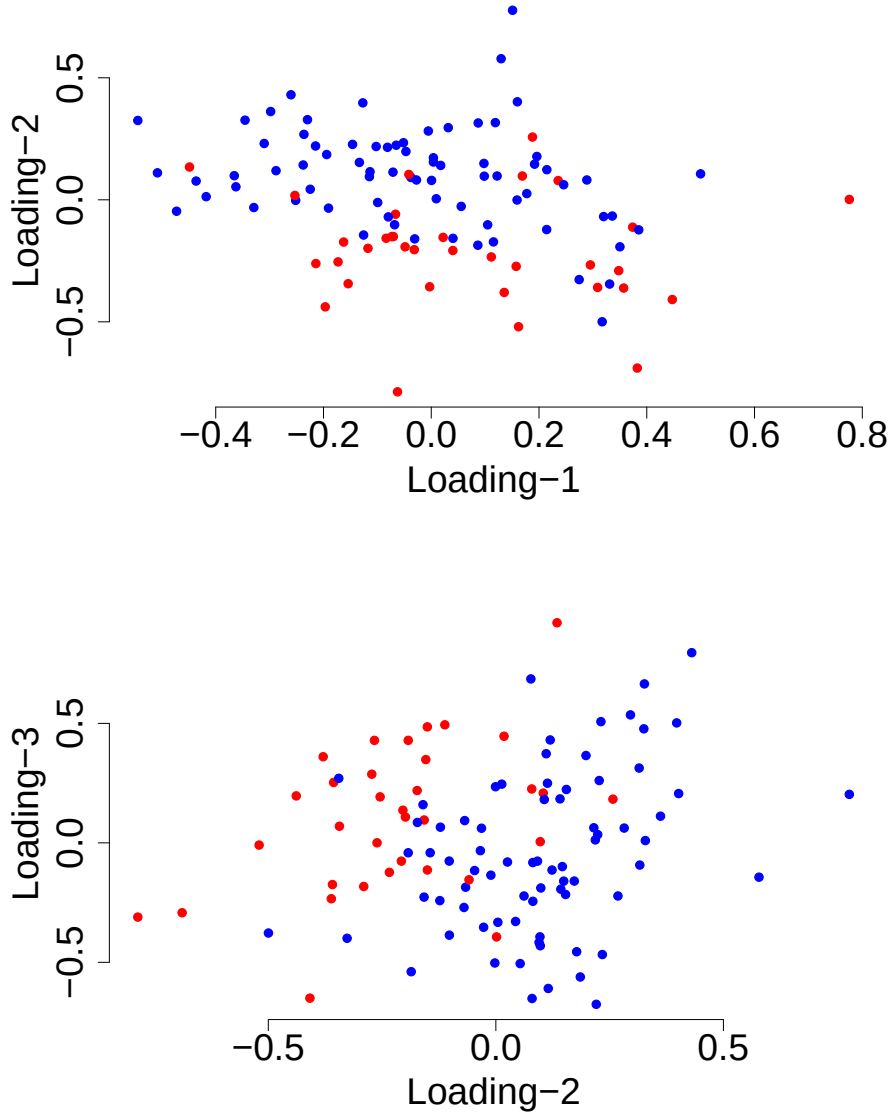

Figure 3.5: Plot showing the separation of smokers (red) and nonsmokers (blue) plotted on the hidden factors learned by HEFT, where the loading 1 and 2 for the factor are plotted on the top and the loading 2 and 3 are plotted at the bottom. The smokers and non-smokers are colored respectively as blue and red.

found 96 non-duplicated significant hits for HEFT using a bonferroni cutoff of $0.05/7575/191959 = 3.438578e - 11$, where the non-duplicated hits are defined

as non-overlapping genes associated with SNPs that are not in linkage disequilibrium. In contrast, the regression analysis returned only 41 associations when using the same criteria. The Quantile-Quantile (QQ) plot of all 1.45e9 p values (Figure 3.6) indicated that while incorporating hidden factors slightly inflated the p value distribution, most of the inflation was in the most extreme p values and the overall plot was well within an acceptable range as measured by a genomic inflation factor of λ<1.07. In addition, a visual inspection of the heat-map of the entire set of p values returned by HEFT compared to linear regression (Figure 3.7), showed that HEFT was able to remove the cases where SNPs were strongly associated with all expressed genes, a clear sign of an unaccounted for nonorthogonal factor [24]. Together, these observations indicate that HEFT is correctly accounting for hidden factors while not over-fitting and, as a consequence, the HEFT analysis is revealing additional eQTL that could not be identified with a naive linear regression. This point is further supported by looking at genes individually, where the HEFT analysis produced well-behaved QQ plots and was also able to reveal significant *cis*- and *trans*- eQTL at a Bonferroni corrected significance threshold (p values < 3.438578e-11) that are not detectable by naive linear regression. Here we present four of these cases (Figure 3.8) that are of particular interest to lung disease, where more are presented in the supplement:

GTF2H1 is the p62 subunit of the multi protein complex transcription factor IIH (TFIIH) that is located on 11p15.1-p14 of chromosome 11. GTF2H1 participates in both the nucleotide excision repair process and transcription control by specifically interacting with a variety of factors important in carcinogenesis. the SNP association we found, rs4150622 is approximately1kb away from a

SNP found to be associated with lung cancer [56]. The second gene, *RUVBL1* is located on 3q21.3 of chromosome 3, is an over expressed gene in several tumors including in the non-small cell lung cancer (NSCLC) tumors [57]. The third gene, TEFM, also known as C17orf42, is located on chromosome 17 that is necessary for transcription of human mtDNA [58]. RNA interference leads to inactivation of TEFM in cells, which leads to respiratory incompetence because of decreased levels of H- and L-strand promoter-distal mitochondrial transcripts. The forth gene, MTRR is located on 5p15.31. Previous studies have shown that variants in or near MTRR show associations with lung cancer in a population of > 2000 non-Hispanic whites[55].

3.5 Conclusion

HEFT is a fast, scalable, and versatile method for detecting eQTL while simultaneously accounting for hidden expression factors that can obscure eQTL signals or produce false positives. Like other hidden factor methods, HEFT is able to avoid false positives that can be produced by hidden factors but the critical advantage of the method is the ability to simultaneously learn hidden factors without over-fitting expression variation. This property results in equal or better performance in comparison to other methods under all conditions, where the advantages are particularly evident when there are pleiotropic eQTL in the presence of hidden factors. The overall advantage of applying HEFT for eQTL detection depends on the existence of potentially problematic hidden factors in a dataset. It does however seems reasonable to assume that hidden factors may be a common problem, particularly when considering cell populations or

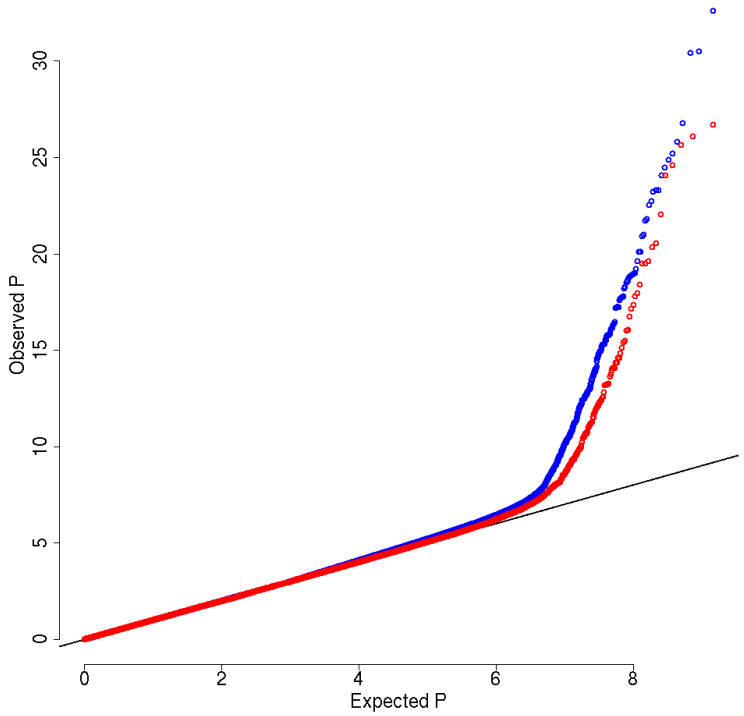

Figure 3.6: QQ plots showing the p value distribution for all tests of association between the 191959 SNPs and 7575 genes expressed in human lung SAE for HEFT (blue) and linear regression (red). Where the quantile of the -Log-P values for the uniform distribution is plotted on the x-axis and the quantile of the -Log-P values for the observed p values are plotted on the y-axis.

tissues collected under uncontrolled conditions, and this assertion supported by our example eQTL analysis, where we were able to find many biologically meaningful eQTL for lung disease using HEFT that are invisible to a standard regression eQTL analysis. We expect the value of methods such as HEFT that can identify additional eQTL while providing strict false positive control will be particularly evident as xQTL studies of new multivariate datatypes available from next-generation sequencing technologies become available for reasonable sized population samples over the next few years. HEFT software that accepts

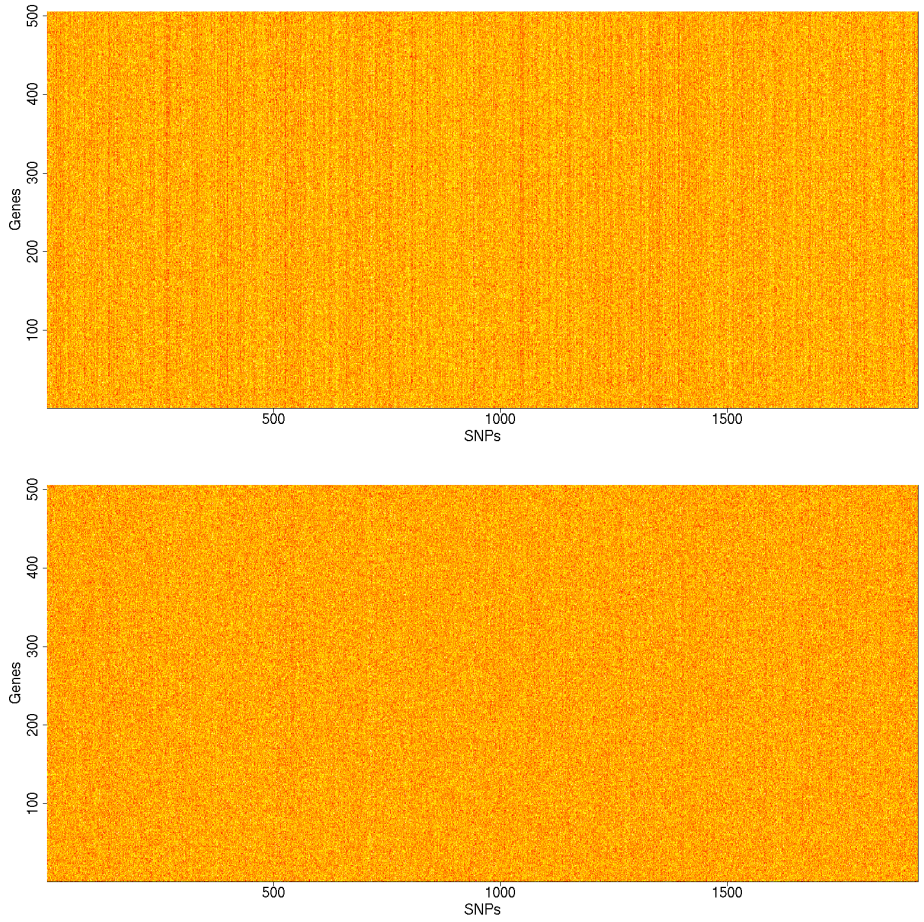

Figure 3.7: A heat map representing p values for tests of association using linear regression (top) and HEFT (bottom) between the 191959 SNPs and 7575 genes expressed in human lung SAE. Genes are arranged in rows and SNPs arranged in columns, where colors transition from yellow to red representing large to small (significant) p values.

standard data formats is available for download, where we plan to incorporate further functionalities such as the ability to analyze genotype-environment interactions and epistasis in the near future.

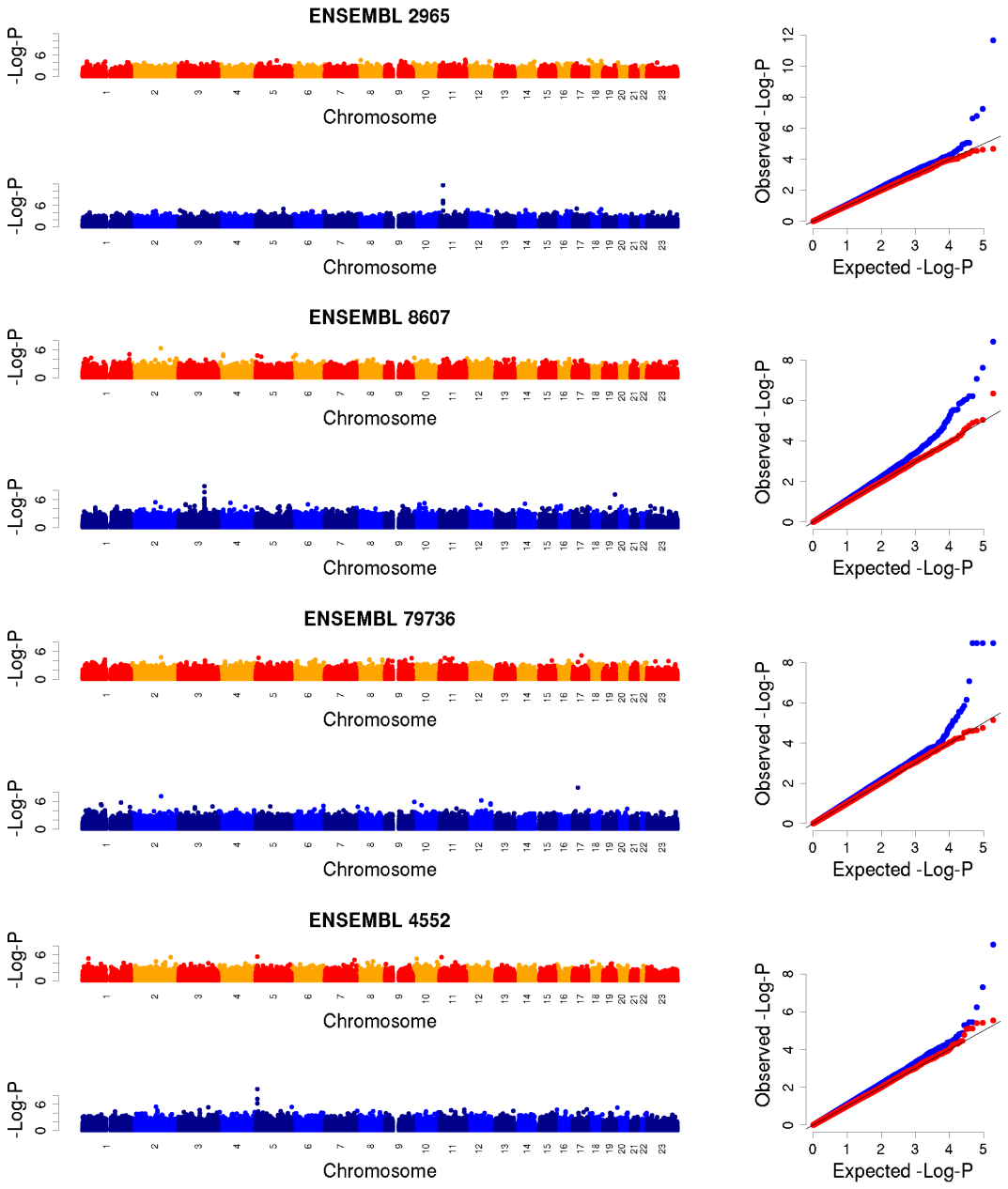

Figure 3.8: Manhattan plots (left column) and QQ plots (right column) for example genes where HEFT (blue plots) identified a significant *cis*-eQTL for a gene with a lung related phenotype or disease association that were not identified by a linear regression (red and orange plots), where the genes ordered from the top to bottom are: GTF2H1, RUVBL1, TEFM, and MTRR.

CHAPTER 4

GENOME-WIDE ANALYSIS OF GENOTYPE-SMOKING INTERACTIONS AFFECTING GENE EXPRESSION IN THE LUNG SMALL AIRWAY EPITHELIUM

4.1 Introduction

Genotype-environment interactions (GEI or GxE), the dependence of genotype-phenotype relationships on environmental factors, is a ubiquitous characteristic of complex traits[72]. In humans, GEI have been found to be important for a variety of complex heritable diseases including coronary heart disease[22], cancer[73], and psychiatric disorders[74]. For many GEI involved in the susceptibility to disease, environmental factors of interest are stress factors such as alcohol consumption[22] or smoking[73, 75], and in many of these cases, the genotype association is only detectable in the presence of the stressor. The increasing number of studies focused on quantifying GEI for disease[76] suggests that the manner in which GEI modulate genotypic effects with environmental stress factors is complex. Understanding the genome-wide extent of GEI is critical, both for the identification of genotype associations and for making accurate predictions of disease risk.

Cigarette smoking is the most important environmental stress relevant to respiratory diseases[73, 75]. Cigarette smoke, with its >103 xenobiotics and 1014 free radicals per puff, places a significant stress on the lung[77, 78, 79], dramatically increasing the risk for bronchogenic carcinoma[80] and chronic obstructive pulmonary disease (COPD)[81]. The cell population most vulnerable to cigarette

smoke is the airway epithelium, the endoderm-derived, pseudostratified layer of cells lining the tracheobronchial tree[82]. The airway epithelium is the first line of defense against cigarette smoke, and it is the epithelium of the small airways that shows the first morphologic changes in smokers[83]. Within the airway epithelium, smoking is known to influence the expression profile of a large number of genes[84, 85] and it is suspected that these changes in gene expression may mediate the effects of smoking on disease[86, 87, 88]. Analysis of smoking effects on genotype-gene expression associations in the airway epithelium is therefore a promising strategy for identifying GEI that may be important for lung disease, and for developing hypotheses concerning the physiological mechanisms responsible for GEI effects on disease.

In the current study we used a genome-wide association (GWA) analysis to analyze GEI on gene expression in the small airway epithelium (SAE) resulting from an individual's smoking environment. Our goal was both to assess GEI genome-wide and to quantify how genetic effects on expression in the SAE can be modulated by smoking. A major difficulty that has hampered the detection and quantification of GEI is that tests of GEI associations are inherently less powerful than tests of the main effect of individual genotypes or the effect of environmental factors considered independently6. Intuitively, the reason for this is that observations are needed for each genotype-environment combination, to produce a test as powerful as a genotype association test.

We employed a combined biological, genomic, and statistical strategy to increase the power of the GEI analysis. The biological strategy was to focus on a tissue (lung SAE) where gene expression is known to be highly responsive to

smoking[84, 21, 89]. Smoking is therefore not only a relevant epidemiological stressor for this tissue, but it is also likely to produce large GEI effects that can be detected with this approach. The genomic strategy was to explore the whole genome by searching all the genotype and gene pairs for GEI, including both *trans* and *cis* effects. Most importantly, we employed a unified hidden expression factor analysis (HEFT) approach to control the hidden confounding in the gene expression data. Previous studies have shown that hidden confounding play a big role in expression Quantitative Trait Loci (eQTL) analysis[21, 53], ignoring these hidden confounding can either lead to false positives or reduce the detecting power of the associations. We hereby used a hidden expression factor analysis approach to control these confounding. HEFT has many nice properties in detecting associations between the multivariate traits and the genotypes, to name a few, 1), HEFT can control for both population structure and gene heterogeneity 2), HEFT does not over correct the genotypic effects, 3), HEFT can incorporate known covariates, 4), HEFT is fast, etc.. These properties make it particularly suitable for detecting GEI effects as it controls for false positives yet without losing true GEI effects.

With this hidden factor approach, we were able to identify significant smoking related GEI on SAE gene expression throughout the genome. The analysis indicates that smoking can change the effect of the genetic determinant dramatically either by enhance it or offset it, especially for those in the top association list, the effect of the genetic variants on its regulated genes can be completely reversed. These GEI relationships add to the complexity of identifying candidate genes associated with smoking-induced lung disease and support the concept that GEI is important for dissecting the impact of abiotic stressors on complex

diseases[90].

4.2 Methods

4.2.1 Study population and sample collection

We used the same 7,575 genes and 191,959 SNPs that we used in chapter 3 for detecting eQTL associations, with the sample demographics shown in table 3.1. For this set of expression traits, t-tests were used to analyze the influence of the following factors: (1) African American vs. European Ancestry; (2) nonsmoking vs. smoking; (3) healthy vs. disease phenotype; and (4) male vs. female. The population stratification were also checked using Principal Components Analysis to eliminating misbehaved samples.

4.2.2 Genome-wide GEI analysis

For the 7575 genes and 191959 SNPs that passed the filter, we used the HEFT model to assess the associations among all gene-SNP pairs. We followed the same model set up with the previous two chapters, that is, we did a single marker analysis by evaluating the genotypic effect one at a time. We included known covariates such as gender, smoking status, disease with each of them being coded up as dummy variables, and the first principal components is included to correct for the population structure. Quite critically, we included the interaction term between the smoking status and the population structure in the model to correct for the false GEI. The full model including the genotypic

effects, the known covariates and the hidden factors can be written as below,

$$\mathbf{Y} = \mu 1'_m + 1_n \beta_0^T + X_A \beta_A^T + X_I \beta_I^T + X_g \beta_g^T + X_s \beta_s^T + X_d \beta_d^T + X_p \beta_p^T + X_{pi} \beta_{pi}^T + \mathbf{\Lambda F} \quad (4.1)$$

where X_A is the additive effects, X_I is the GEI, X_g is the gender effect, X_s is the smoking effect, X_d is the disease status, X_p is the first principal component of the genotype and X_{pi} is the interaction between the smoking and the population structure. All other parameters are the same as in chapter 3. Considering that we have included several known covariates, We again conservatively selected 5 factors based the eigen spectrum of the gene expression matrix to avoid over fitting the model.

For the genome-wide GEI analysis, we used the false discovery rate (FDR) to adjust for multiple tests, when determining significant genotype-smoking interaction (GEI). The most significant GEI associations were identified at an experiment-wide false discovery rate (FDR) calculated similarly to the Benjamin and Hochberg procedure[91]. The Benjamin and Hochberg is calculated as following, for a number of p values calculated from q test, $p_1 \leq p_2 \leq ... \leq p_q$, and let α be the desired significance level, the procedure compares p_j with $\alpha j/q$ until $p_j \leq \alpha j/q$, The benjamin Hochberg procedure goes through this list from the most insignificant p values to the most significant ones. However, since in our case, we have an extremely long list of non-significant p values, I proceeded in the reverse order, that is, starting the search backward from the most significant to the most insignificant ones until $p_j \geq \alpha j/q$, which yield a more conservative cutoff. To avoid either overfitting or under fitting the model, the t statistic was adjusted by using a genomic control approach before being ranked. Specifically, for the p values across the SNPs for a specific gene, a inflation factor λ was calculated by taking the raio of the median of observed the t statistic and the

median from the expected t distribution, then the t statistic for this gene across the genome were adjusted by this value.

4.3 Results

4.3.1 Gene expression in the small airway epithelium

The gene expression analysis on the 7,575 genes shows that smoking had a large effect on the overall expression profile, resulting in a greater than 1.3-fold up-regulation of 12 genes and down-regulation of 0 genes (Fig. 4.1). The genes with the most significant and largest smoking effects overall were cases of up-regulation. For significant cases as determined by a t-test using a Bonferroni correction $p < (0.05/7575) = 6.6E\text{-}6$ the ratio of up-regulated genes to down-regulated genes was 267/209=1.28. Lung disease, and ancestry had a far more limited effect on overall SAE expression than smoking, with very few differences >1.3-fold. However, gender does show dramatic effects on a limited number of genes, although the number of genes affected by gender don't compare to those by the smoking status, the most significant ones show far more significant p values, which indicate genes expressed differently in males and females are very distinct. (Fig. 4.1).

4.3.2 Genotype and smoking interaction across the genome

After running HEFT on all 7575 genes and 191959 SNPs and the p value matrix were obtained, we first adjust the p values for each gene across the whole

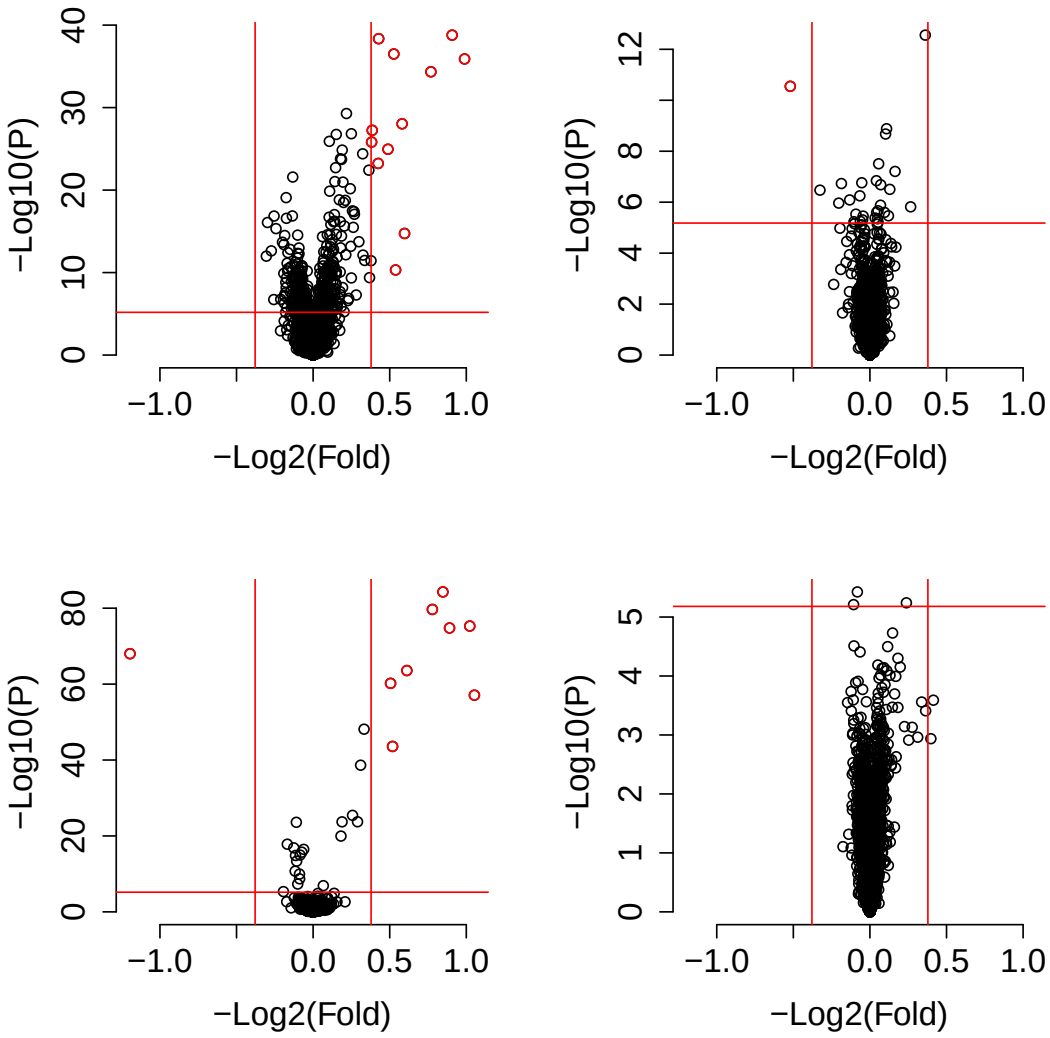

Figure 4.1: Gene expression in the small airway epithelium (SAE). Shown clockwise are volcano plots of the effects of (a) smoking; (b) ethnicity; (c) gender; and (d) disease status on SAE gene expression. The vertical red lines correspond to log2 transformed fold change of 1.3 and 1/1.3. The horizontal red line corresponds to the Bonferroni corrected significance threshold p = 0.05 for individual t-tests performed on each gene. Effects beyond these thresholds for both fold change and *p* value are shown in red

genome by using the genomic control approach mentioned in the method section, this gives a fair comparison of all the p values, as overfitting the model by using more factors can lead to inflated p values, and under fitting by using less factors lead to deflated p values. The genomic control on the p values for all gene-genotype pairs bring them onto the same baseline, so that no p values will be ranked on the top due inflation, and vice versa for the deflation. After ranking the p values, we found that no associations passed the extreme bonferroni correction criteria which is calculated at 0.05/7575/191959 = 3.438578e-11 due to the inherent low power of the GEI test, especially with the small sample size. Using the Benjamin and Hochberg procedure FDR procedure, I get thousands of associations based on a FDR cutoff 5e-6

Many of the genes exhibiting significant GEI are of interest in terms of lung biology and smoking-induced lung disease, we identified a large number of interesting GEIs that are for biological relevant genes that have been verified to play important role in lung function, lung disease or functions related to smoking. we selected 2 interesting genes to show in the main text and the rest can be found the supplementary materials. The first GEI we found particularly interesting is for gene TLR4 (Figure 4.2), the toll-like receptor 4. This gene is located on chromosome 9q33.1, it encode proteins that is a member of the Toll-like receptor (TLR) family, which plays a fundamental role in pathogen recognition and activation of innate immunity. study has shown that TLR4 expressed in human lung cancer cells is functionally active, and may play important roles in promoting immune escape of human lung cancer cells by inducing immunosuppressive cytokines and apoptosis resistance[92]. The most significant genetic variant I found, rs322006 is located in the intron region of the

gene membrane associated guanylate kinase, WW and PDZ domain containing 2 (MAGI2), it turns out that MAGI-2 enhances the ability of PTEN to suppress Akt activation[93], and PTEN supports TLR4-induced inflammatory responses by suppressing the activation of Akt[94]. This GEI finding seemed to suggest that MAGIC2 regulate the TLR4 differently in smokers and non-smokers. Since smokers is more easily to get cancer, A bold guess would be MAGI2 is expressed higher in smokers than non-smokers, which enhanced the ability of PTEN to suppress Akt, this subsequently increased the TLR4 signaling, helping the lung cancer cell to escape the immune system by inducing immunosuppressive cytokines and apoptosis resistance.

The second interesting GEI example we found is for gene SIN3 transcription regulator homolog A (SIN3A) (Figure 4.3), which is located on Chromosome 15q24.2, previous study have shown that the attenuated function of SIN3A due to a decreased level of expression may result in epigenetic de-regulation of growth-related genes through histone acetylation, which leads to the tumorigenesis of lung cancer cells[95]. We found a genetic variant rs2325834 for it, which is located in the intron region of gene CDH13. Interestingly, studies have shown that a combination of deletion and hyper methylation causes inactivation of the CDH13 gene in a considerable number of human lung cancers as well[96]. Combining the evidence from previous studies and the Interaction plot shown in Figure 4.3. It seemed like that the gene SIN3A act as a regulator of CDH13 which behaves differently in smokers and non-smokers. eg., for non-smokers, the expression level for SIN3A increase as certain copy of the allele increase, which up regulate CDH13 and suppress the tumorigenesis of the lung cancer cell. However, in smokers, the trend is exactly reversed, leading to the growth

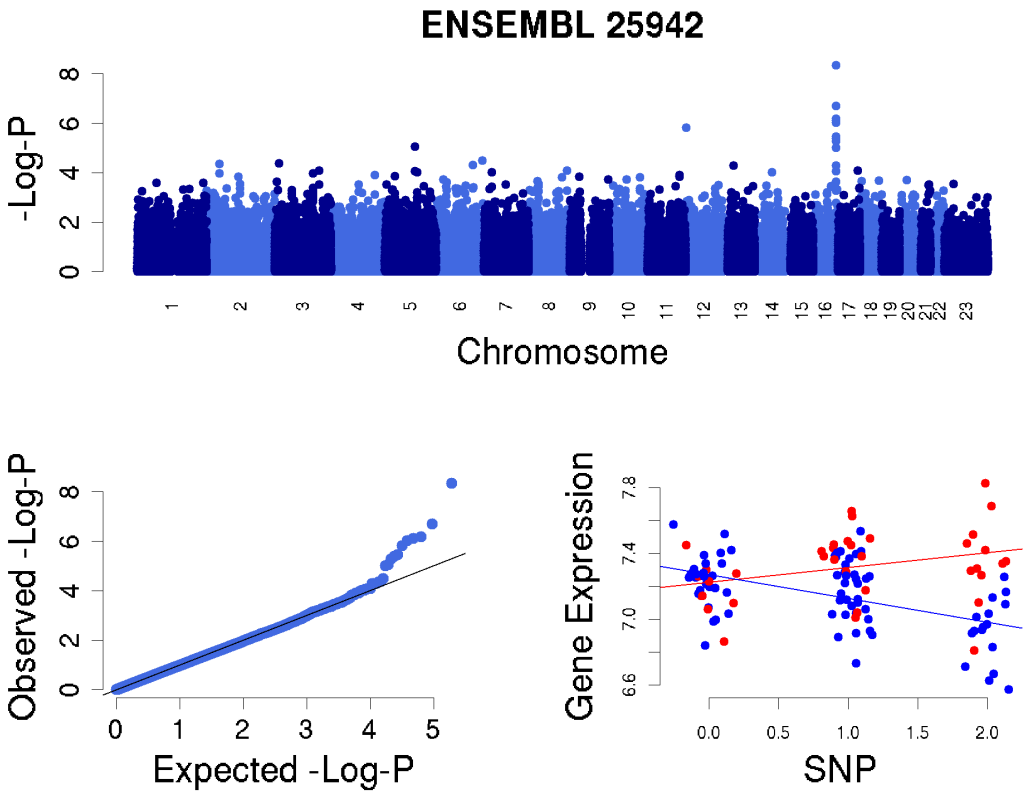

Figure 4.2: The Manhattan plot, QQ plot for the p values of gene TRL4 and its associations with SNPs genome wide. The scatter plot of the gene expression measurement and the associated SNPs showing the interaction between the two are also shown, where manhattan is plotted on the top, QQ is the bottom left and GEI scatter plot is on the bottom right. For the GEI plot, red and blue color correspond to non-smoker and smokers respectively.

of the tumor cells.

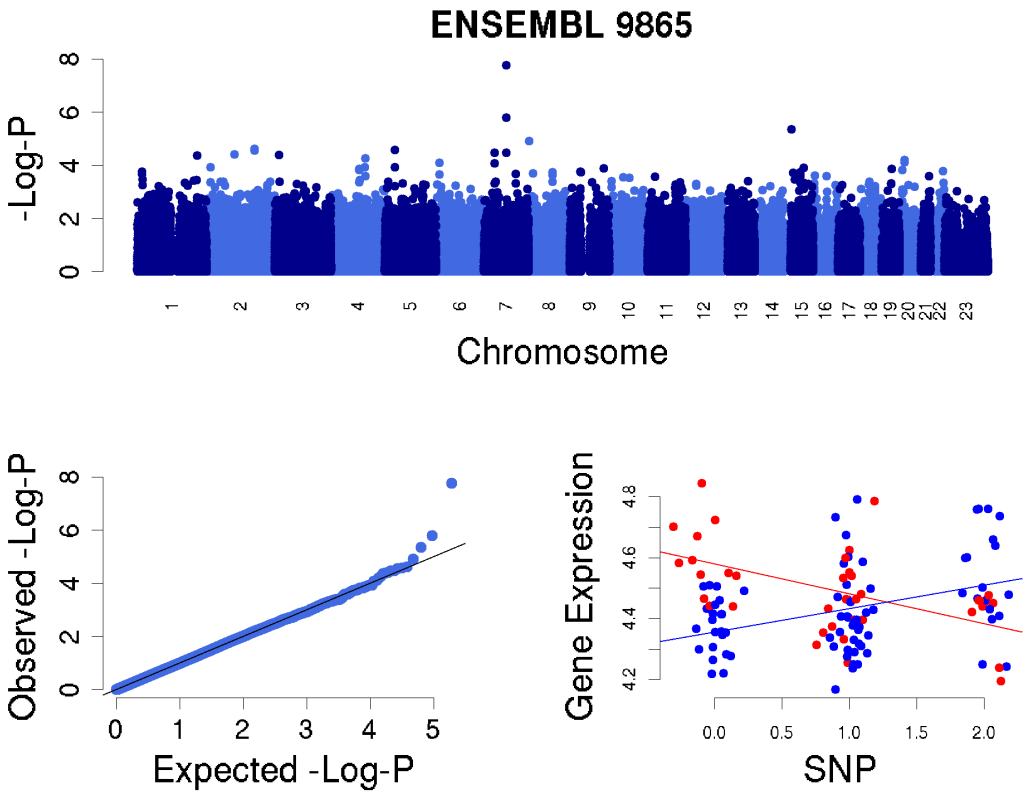

Figure 4.3: The Manhattan plot, QQ plot for the p values of gene SIN3A and its associations with SNPs genome wide. The scatter plot of the gene expression measurement and the associated SNPs showing the interaction between the two are also shown, where manhattan is plotted on the top, QQ is the bottom left and GEI scatter plot is on the bottom right. For the GEI plot, red and blue color correspond to non-smoker and smokers respectively.

4.4 Discussion

The current study is one of the first genome-wide analyses of gene expression GEI in humans, given the small sample size and the limited power to detect GEI, the large number of GEI that we identified are likely to reflect a considerable underestimate of the total influence of GEI on SAE expression. The analysis

therefore indicates that GEI has a considerable influence on the relationships between genotype and gene expression in the lung small airway epithelium.

Given that smoking places a considerable stress on the small airway epithelium of the lung and causes a dramatic alteration of gene expression genome-wide in this cell population[89, 84, 21], we might expect the GEI identified in this study to be more extreme than GEI associated with other environmental agents. However, given that as of 2007, 19.8% of the US population are smokers and a far larger percentage encounter second hand smoke[97], these results at the very least indicate that GEI are critical for understanding heritable components of global gene expression variation in the lung. It is also possible that these results for smoking are indicative of extensive GEI associated with environmental stressors on specific cell populations. The result of this study may therefore be indicative of a broader trend whereby abiotic factors, such as pollution and other xenobiotics, tend to modulate genotype effects on gene expression. Since relative expression levels of many genes are expected to provide an indicator of disease risk, and in many cases may be directly responsible for disease, our analysis indicates that GEI are a critical component of the connection between genotype, gene expression, and complex diseases.

CHAPTER 5

SUPPLEMENTARY MATERIALS

5.1 Supplementary tables for HEFT

Table 5.1: Table showing the parameter set up for all simulations to show the combinations of hidden factors, eQTLs and Pleiotropic eQTLs, where × says that parameters are absent and √ indicate the parameters are present. All simulated parameters are similar for all cases. Please see the method section for more detail. The bottom rows show the heritability range for the all simulated scenarios for both non-orthogonal factors and orthogonal factors.

	Scenarios					
	a	b	c	d	e	f
eQTLs	×	×	√	√	√	√
Pleiotropic eQTLs	×	×	×	√	×	√
Factors	×	√	×	×	√	√
Heritability(min/max)-non-orth	×	×	3.6e-06/0.81	8.6e-07/0.91	2.1e-06/0.81	5.5e-07/0.87
Heritability(min/max)-orth	×	×	3.9e-06/0.84	7.2e-07/0.92	1.1e-06/0.84	5.0e-07/0.83

Table 5.3: table of 100 non-duplicated top GEI associations identified by HEFT, where only the top associated SNPs were liste. From left to right the columns represent respectively the ranking, ensemble code for the genes, the top SNP associated with this gene, and the p values. The whole list is available per request.

1	5906	1	112038561	rs1886498	1.46499485948337e-33
2	140686	20	43836276	rs2664529	2.27127502421512e-31

3	89778	18	59530818	rs4940595	7.75169142022207e-28
4	374491	13	24078151	rs943049	1.50554696527431e-25
5	80177	6	153052613	rs2250514	1.74101297874152e-25
6	6840	0	0	rs3858231	2.12530111251114e-25
7	63928	16	23697839	rs194788	8.46873720736218e-25
8	23421	1	63837368	rs855325	3.53979830579298e-24
9	114757	17	72058534	rs752049	7.33012802900279e-24
10	388335	17	10556076	rs397278	6.39394600924099e-23
11	1915	6	74287580	rs3822960	7.95367824562483e-22
12	5340	6	161064326	rs1321200	6.63854323347397e-21
13	116285	16	20587707	rs433598	2.02376243218272e-20
14	155368	7	72863628	rs4355658	2.35431495370645e-20
15	340542	23	101168380	rs2858353	2.61556705576623e-20
16	8000	8	143761003	rs2976396	5.75349616679876e-20
17	164781	2	228485182	rs3748863	6.47663988110386e-20
18	51144	11	43796511	rs10768983	9.78888533972432e-20
19	318	9	34271390	rs7045680	1.26766986219884e-19
20	5947	3	140736561	rs12485273	1.86700120545829e-19
21	26751	2	270819	rs7605824	4.43726175926121e-19
22	403314	1	181889002	rs6699011	1.26809312494076e-18
23	7976	8	28491587	rs11779401	1.77971301988934e-18
24	90637	7	1171226	rs2960840	2.07643459155093e-18
25	25961	10	74540513	rs2280369	5.3059275592548e-18
26	55278	6	107221054	rs1026619	1.17621599828576e-17
27	1965	14	66916971	rs8008724	1.33675165066165e-17
28	150142	21	42317483	rs220219	1.64955608207078e-17

29	286464	23	36068724	rs16987374	1.77494728717695e-17
30	22948	5	10318076	rs699113	1.86493789973556e-17
31	158158	9	84807488	rs1502682	3.86943395919504e-17
32	5268	18	59295033	rs3744941	5.34526015688907e-17
33	26503	6	74399417	rs9446964	8.09618716653111e-17
34	26090	20	25223843	rs2258719	1.45642163092577e-16
35	54879	1	112933916	rs6666579	1.66651287593059e-16
36	10230	17	38773860	rs4793229	1.72080169717462e-16
37	100506015	2	161907239	rs10197817	2.54805768511242e-16
38	100507580	22	24215353	rs6004673	4.35569747641881e-16
39	91612	14	64477410	rs2412065	4.44742133568217e-16
40	84930	10	27494117	rs7068375	7.17212767471488e-16
41	4649	15	69903832	rs2742323	1.59951937168628e-15
42	84545	10	102724768	rs4919510	2.19436097956699e-15
43	10781	19	9394805	rs6512121	2.6373840456275e-15
44	388407	17	56851431	rs2079795	2.91445531827876e-15
45	5889	17	54195586	rs8074016	3.72419479226112e-15
46	25854	4	187345974	rs4586997	3.77686506053501e-15
47	80150	11	61840106	rs1406384	7.16689214689067e-15
48	100131564	1	93561736	rs7555292	1.02301175982386e-14
49	55034	18	31980483	rs3737468	2.41283860647574e-14
50	121506	12	14961902	rs2193356	2.68918264415563e-14
51	26999	5	156687851	rs13155266	4.73999050614003e-14
52	100507540	9	74103814	rs7874628	4.77268615823305e-14
53	10558	9	93804054	rs7045602	5.39021021351054e-14
54	51531	9	99670770	rs7357707	5.90037022855741e-14

55	2882	1	52836600	rs835341	6.04084235032921e-14
56	54960	23	13929093	rs7055913	7.41803216920483e-14
57	27030	14	74615438	rs175490	9.02909469073108e-14
58	51703	10	114138679	rs12255316	1.16488754370987e-13
59	55020	22	45066872	rs6008552	1.42090376672065e-13
60	538	23	77217818	rs2643591	1.58701624661496e-13
61	146562	16	4734874	rs2075469	1.70605590922464e-13
62	6263	15	31881493	rs2115747	2.80777022150772e-13
63	100507316	8	144410365	rs7824894	2.90777614923962e-13
64	51816	22	16062294	rs1076106	3.04711413440298e-13
65	8624	21	39555501	rs2836965	3.54602705122609e-13
66	143241	10	82062290	rs10788562	4.6114142614977e-13
67	3631	2	98402962	rs17031139	5.2297407172801e-13
68	51016	14	23681818	rs3742500	5.62602288154044e-13
69	622	3	198745276	rs13077136	1.03436568736022e-12
70	201651	3	152973681	rs4679934	1.28939628281059e-12
71	133383	5	56214969	rs832584	1.92293301156847e-12
72	2965	11	18322286	rs4150622	2.14534949336851e-12
73	55253	7	66219599	rs17144722	2.52141749626425e-12
74	84221	21	46524284	rs2839195	2.64339920859971e-12
75	151525	2	159886945	rs174227	3.8214945233408e-12
76	643529	10	91582178	rs1125326	4.19750709824653e-12
77	64105	5	64951379	rs2161278	4.27951343576203e-12
78	654433	2	113700504	rs3748916	4.56219494155033e-12
79	100506707	2	113705738	rs2863243	6.6617384815837e-12
80	55728	4	39796815	rs17619330	6.95256644737194e-12

81	401491	9	2527815	rs588933	8.40770283486909e-12
82	80868	6	30026415	rs2508037	8.68017934939448e-12
83	7180	6	49811816	rs597544	1.10976804567039e-11
84	55125	18	12967206	rs8088313	1.21441064557634e-11
85	54847	3	114737833	rs4580515	1.24583836679942e-11
86	57545	4	15091907	rs6810461	1.45170694299707e-11
87	11102	3	58282684	rs6777105	1.45641838150676e-11
88	284323	19	45181310	rs8105066	1.52901564045426e-11
89	10783	9	126069557	rs12379417	1.59002247514929e-11
90	55256	2	3488694	rs9750132	1.62436743738252e-11
91	145957	15	73938020	rs2593280	1.89118641889377e-11
92	6006	1	25641524	rs10903129	2.2380711377044e-11
93	55733	1	208584705	rs6696657	2.34656975928403e-11
94	128344	1	111697010	rs1058530	2.58389549266189e-11
95	79772	5	93992505	rs10052066	2.95141473257443e-11
96	79618	8	28881393	rs4732896	3.24184877322828e-11
97	55234	9	33035161	rs10758181	3.53463603710357e-11
98	28512	3	23965245	rs11922577	3.59932003158567e-11
99	286042	8	8185680	rs2945886	3.62424068477289e-11
100	57115	1	151585677	rs2916212	3.81258886782951e-11

5.2 Supplementary Figures for HEFT

5.3 Supplementary tables for GEI

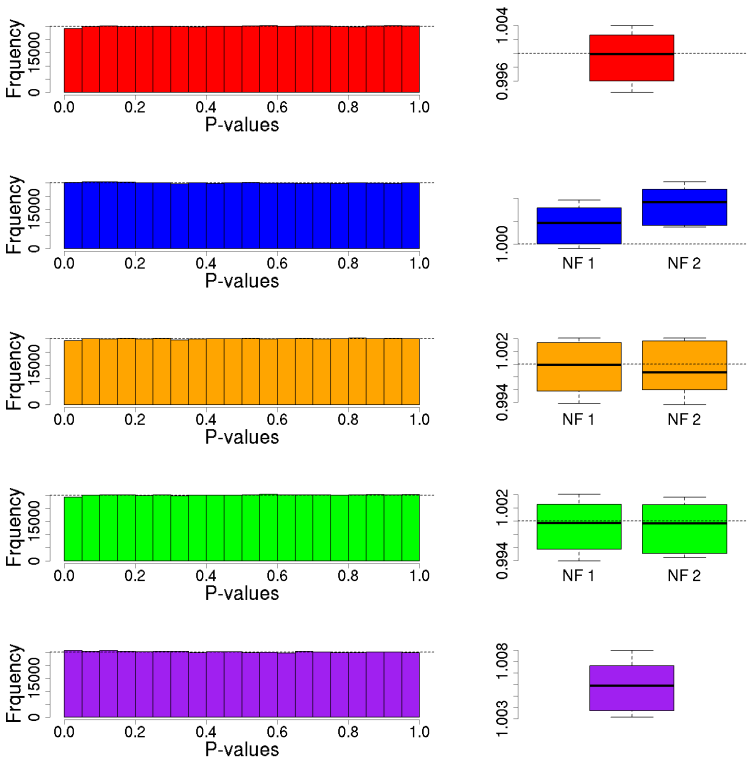

Figure 5.1: Histograms and boxplots showing the distributions of p value for all SNP-gene tests of association for the scenario where there are no genotypic effects and no hidden factors (scenario a). The left column shows the histogram of the p values for a specific simulation with factor number of 2 (when the factor number applies), and the right column shows the boxplots of the inflation factor for p values of all ten simulations. From top to bottom are respectively linear regression, HEFT, HEFT-TS, VBQTL and LMM-EH. For the boxplot, from left to right corresponds to factor number of 1 and 2 when factor numbers apply

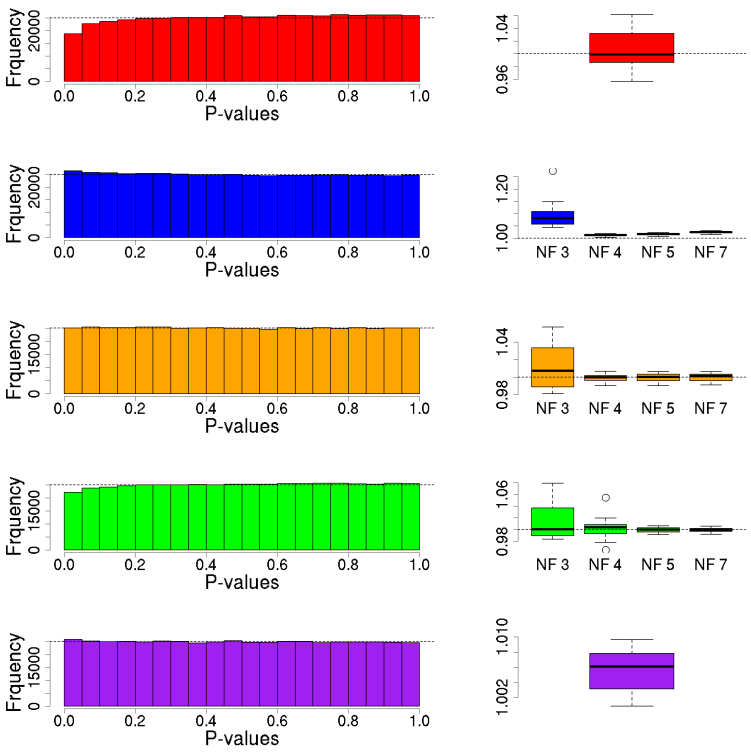

Figure 5.2: Histograms and boxplots showing the distributions of p value for all SNP-gene tests of association for a scenario with no eQTL and hidden factors that are orthogonal to the SNPs(orthogonal scenario b). The left column shows the histogram of the p values for a specific simulation with factor number of 7 (when the factor number applies), and the right column shows the boxplots of the inflation factor for p values of all ten simulations. From top to bottom are respectively linear regression, HEFT, HEFT-TS, VBQTL and LMM-EH. For the boxplot, from left to right corresponds to factor number of 3,4,5,7 when factor numbers apply.

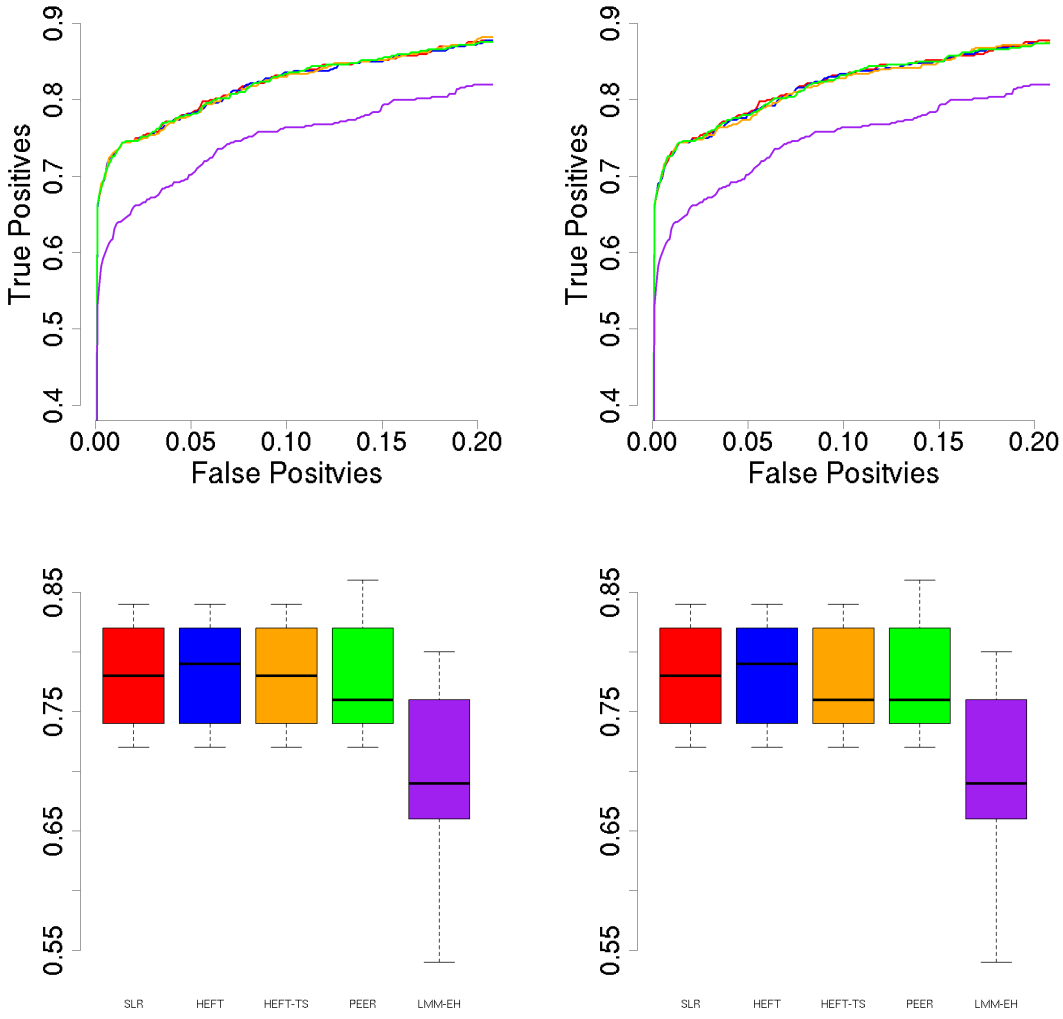

Figure 5.3: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenarios where there are eQTLs but no hidden factors, where the left and right columns correspond to provided factor numbers of 1 and 2. The methods are color coded as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

Table 5.2: Table showing the inflation factor for all methods on scenario a and b, where a plot of a subset of these can be found in the main text and the supplementary figures. the second row of the table shows the factor number used for methods that require such a number. And the values in the table shows the confidence interval of the inflation factor for all 10 simulations of each scenario.

		a		b			
	#Factor	1	2	3	4	5	6
non-orthogonal	SLR	1±0	1±0	1.21±0.04	1.21±0.04	1.21±0.04	1.21±0.04
	HEFT	1±0	1.01±0	1.33±0.12	1.01±0	1.02±0	1.02±0
	HEFT-TS	1±0	1±0	1.14±0.1	0.99±0	0.99±0	0.99±0
	VBQTL	1±0	1±0	1.16±0.08	1.1±0.08	0.99±0	0.99±0
	LMM-EH	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0
orthogonal	SLR	1±0	1±0	1±0.03	1±0.03	1±0.03	1±0.03
	HEFT	1±0	1.01±0	1.1±0.07	1.01±0	1.02±0	1.02±0
	HEFT-TS	1±0	1±0	1.01±0.02	1±0	1±0	1±0
	VBQTL	1±0	1±0	1.01±0.03	1±0.02	1±0	1±0
	LMM-EH	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0	1.01±0

Table 5.4: table of 100 non-duplicated top GEI associations identified by HEFT, where only the top associated SNPs were liste. From left to right the columns represent respectively the ranking, ensemble code for the genes, the top SNP associated with this gene, and the p values. The whole list is available per request.

Ranking	Gene	Chr	Position	SNP	P value
1	390940	10	61187929	rs1913509	2.12252512341375e-10
2	7503	6	42298534	rs4714584	2.55999326509644e-09
3	56970	11	95457552	rs1939478	4.24641470800478e-09

4	25942	16	82057455	rs2325834	4.52986233115035e-09
5	64949	16	26777178	rs2188776	5.58942686820258e-09
6	83590	1	86671145	rs1931363	5.76960485893903e-09
7	63929	8	19326299	rs6586829	9.42718246981585e-09
8	8763	17	73423766	rs16970654	9.66716995052864e-09
9	158401	7	33958400	rs10270579	1.02036772837876e-08
10	5863	3	63050307	rs1374679	1.20520296508348e-08
11	118424	10	97122546	rs11188298	1.54580331233261e-08
12	170425	1	72619048	rs1599337	1.59879098209009e-08
13	6263	5	23072125	rs12657604	1.6408856181601e-08
14	100506802	3	30593715	rs9815023	1.65553600401073e-08
15	9865	7	78117685	rs322006	1.69970027592e-08
16	84996	14	76005578	rs9944093	1.71613357302719e-08
17	677	11	123320365	rs4300403	1.73416414199539e-08
18	7073	17	45394124	rs1017285	2.07337854656208e-08
19	153768	20	47960830	rs2769978	2.16357849708456e-08
20	54830	7	36248233	rs4723501	2.17521048163392e-08
21	81792	1	174755116	rs10127938	2.37199032574018e-08
22	4234	3	1870810	rs2029359	2.56386451975302e-08
23	60437	4	99187100	rs7666738	2.59088333308738e-08
24	90506	5	171980753	rs7704716	2.73072847305238e-08
25	150350	11	81286656	rs1945897	3.12231285260636e-08
26	23139	5	6378444	rs13177370	3.2959736729877e-08
27	114571	1	30832451	rs555920	3.47701008312964e-08
28	4855	2	84716354	rs4313996	3.49676293880892e-08
29	3069	18	13794755	rs2105655	3.52308650191333e-08

30	56955	5	155577369	rs249879	3.71196679797979e-08
31	23612	10	133004199	rs7080842	3.79603276429811e-08
32	256472	3	59944934	rs3772492	3.97327943055445e-08
33	158067	7	41771061	rs17638578	4.18242509036606e-08
34	4172	8	8510570	rs11784888	4.2808908783042e-08
35	442075	2	56595255	rs7605943	4.28505872917591e-08
36	6050	4	13286139	rs966385	4.44221546287363e-08
37	100506124	13	94441938	rs9516507	4.45592133386073e-08
38	84457	5	168576631	rs8180402	4.75463857226505e-08
39	400360	3	177487380	rs16826837	5.24270286860765e-08
40	29924	5	153546572	rs7719182	5.25915691439414e-08
41	5709	8	104123779	rs7009365	5.51953153704643e-08
42	7433	11	55264743	rs10897182	5.52418172027518e-08
43	161835	3	115577011	rs870248	5.59402738796233e-08
44	89885	18	13370824	rs1026330	5.61912633073283e-08
45	439921	2	224569405	rs11695803	5.76589167349634e-08
46	399979	9	102943652	rs2567305	5.82757506017848e-08
47	22983	8	39092304	rs4733975	5.87267956574457e-08
48	3626	9	27342095	rs7035694	5.9871296391004e-08
49	784	11	68488253	rs4244842	6.09997019728339e-08
50	27239	14	67078620	rs3742867	6.13398286028766e-08
51	416	11	9753275	rs10770061	6.27748853761829e-08
52	25992	18	42864524	rs16954921	6.42426989821352e-08
53	6452	14	88634883	rs4904503	6.77210095848897e-08
54	9093	23	132993225	rs6638155	7.19635032576722e-08
55	53905	2	86238746	rs7561589	7.27351844582759e-08

56	337	7	18259746	rs302137	7.42764695204938e-08
57	93650	12	19927556	rs16915343	7.56622821960498e-08
58	85364	11	19477837	rs7103939	7.65861182958018e-08
59	51237	3	13967421	rs4685052	8.28952105969382e-08
60	10943	5	175934129	rs10476198	8.36909742831747e-08
61	100505565	1	112263425	rs1443926	8.42820761205385e-08
62	9828	3	2810174	rs9847255	8.74274523410288e-08
63	54914	2	118808814	rs13426749	8.76808121691761e-08
64	221002	8	34348223	rs978439	8.81939079785246e-08
65	139135	12	66025356	rs11176693	8.99493409220536e-08
66	51330	16	62977010	rs918732	9.60303920288732e-08
67	119	3	151317695	rs6785242	9.66232293921189e-08
68	5522	16	81868479	rs2042434	9.72273647420355e-08
69	440184	11	36726968	rs11033790	9.84784488060625e-08
70	9380	6	119874238	rs6936821	1.02690835592991e-07
71	3485	11	33204872	rs11606914	1.03135291954957e-07
72	377677	8	30471800	rs7812836	1.05659860586791e-07
73	85021	1	80927938	rs17105403	1.05887797636833e-07
74	51204	5	169970289	rs1013922	1.06161735132535e-07
75	283401	16	7309267	rs8046170	1.06389643851991e-07
76	150	23	40158580	rs2056491	1.07006673671982e-07
77	283987	10	127910717	rs3858313	1.07816414170571e-07
78	25851	5	119396967	rs7722124	1.10060414941561e-07
79	2009	2	230965056	rs7559665	1.10083185248058e-07
80	79581	1	204036174	rs2224	1.11539665357704e-07
81	58513	10	43814552	rs10793513	1.11610347139357e-07

82	285051	18	59771746	rs9945924	1.119725702033e-07
83	1585	17	5976209	rs4426395	1.15035573026416e-07
84	10466	18	59875096	rs7244757	1.16221382326436e-07
85	284454	22	44968942	rs12330015	1.16977094139164e-07
86	7957	6	97360457	rs9487199	1.24015621365315e-07
87	8884	18	55493989	rs10503031	1.25621355711241e-07
88	55180	13	19876006	rs7165	1.25883976398338e-07
89	23131	8	10370887	rs7833781	1.26343725182718e-07
90	11009	7	127645127	rs2167289	1.27229986365669e-07
91	3099	8	4436914	rs17070710	1.29015511076561e-07
92	728489	3	96350425	rs9867698	1.29818852036134e-07
93	6156	20	921669	rs1569816	1.32837914285109e-07
94	9918	23	127037573	rs1585268	1.34780160094746e-07
95	54093	13	52496541	rs9536340	1.3543393711456e-07
96	5625	9	1052676	rs10959157	1.36897264084116e-07
97	2242	1	4714260	rs428001	1.37235593789223e-07
98	585	15	77929052	rs655473	1.38202010143116e-07
99	414	6	36033776	rs10456081	1.38753054854132e-07
100	126074	4	14815436	rs10516284	1.3939663283278e-07

5.4 Supplementary Figures for GEI

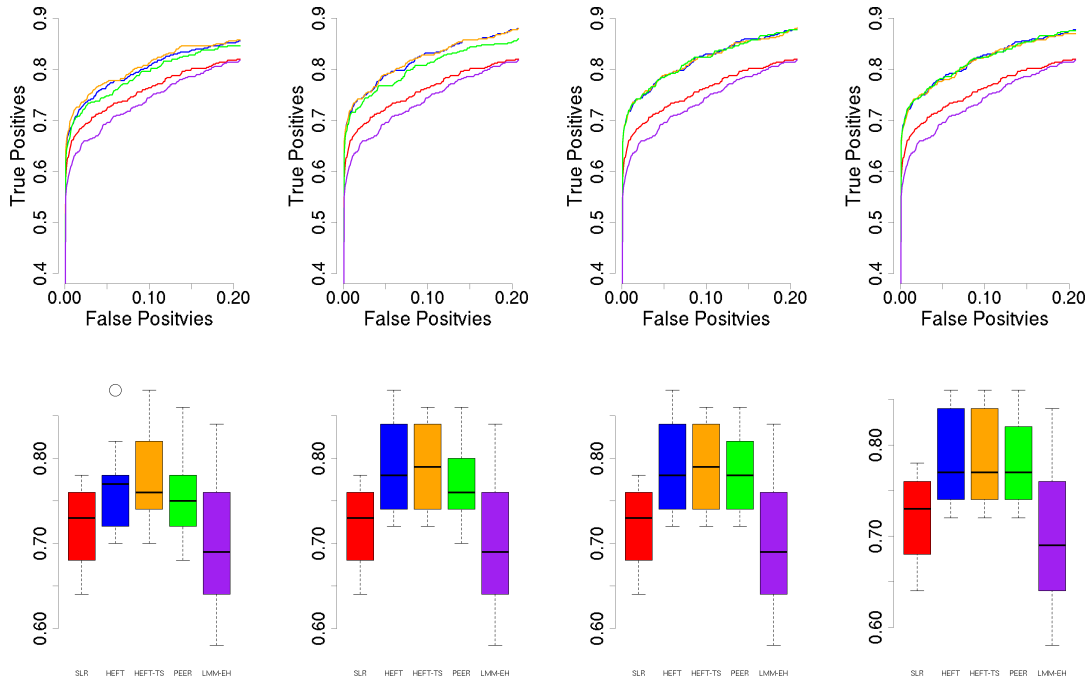

Figure 5.4: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenarios where there are eQTL effects and orthogonal hidden factors (scenario e), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively. The methods are color coded as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

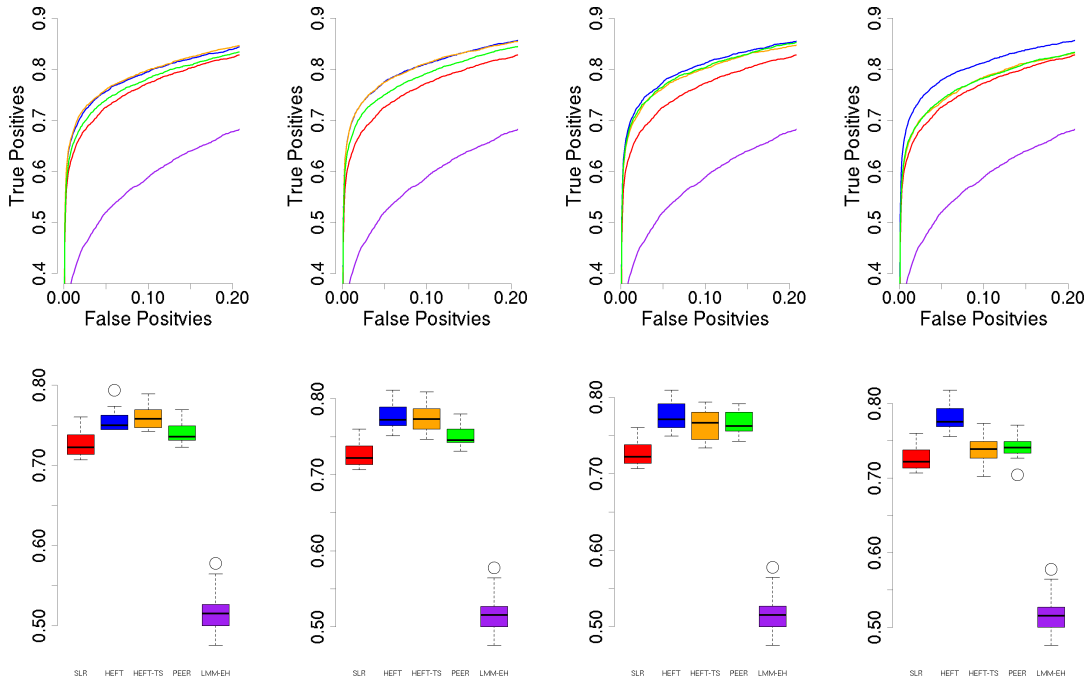

Figure 5.5: Receiver Operative Characteristic (ROC) curves (top) and boxplots of the true positives at false positive rate of 0.05 (bottom) for all five methods for scenarios where there are pleiotropic eQTL effects and orthogonal hidden factors (scenario f), where from left to right correspond to provided factor numbers of 3, 4, 5, and 7 respectively. The methods are color coded as in figure 3.1 (red=regression, blue=HEFT, orange=HEFT-TS. green=VBQTL, purple=LMM-EH.)

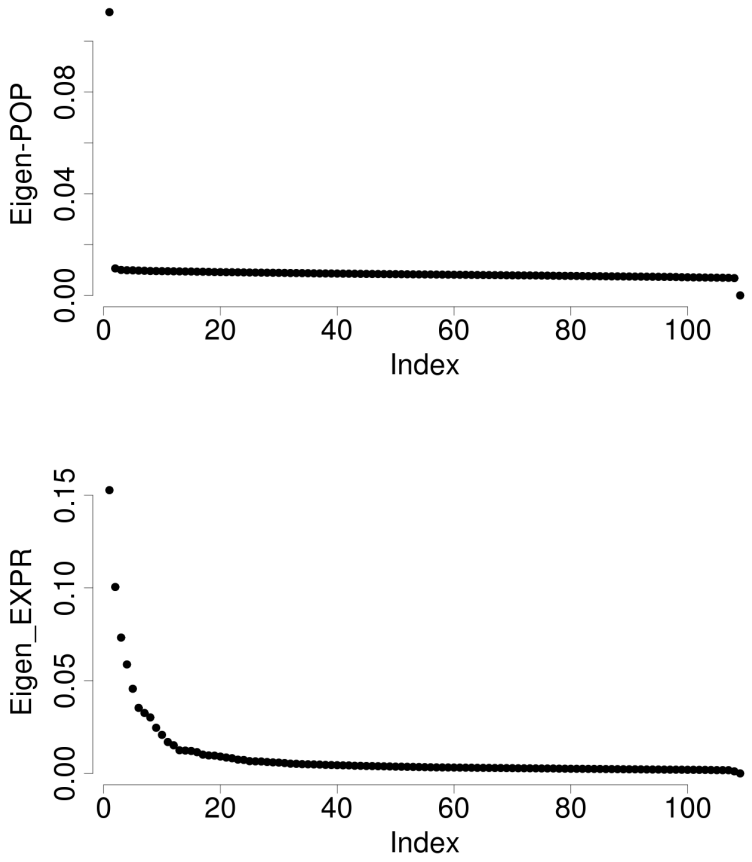

Figure 5.6: Eigen spectrum plot for the genotypes (top) and the gene expression data (bottom), where the x-axis shows the index of the eigen values (eigen vectors) and the y-axis shows the variance proportion explained by each eigen vector.

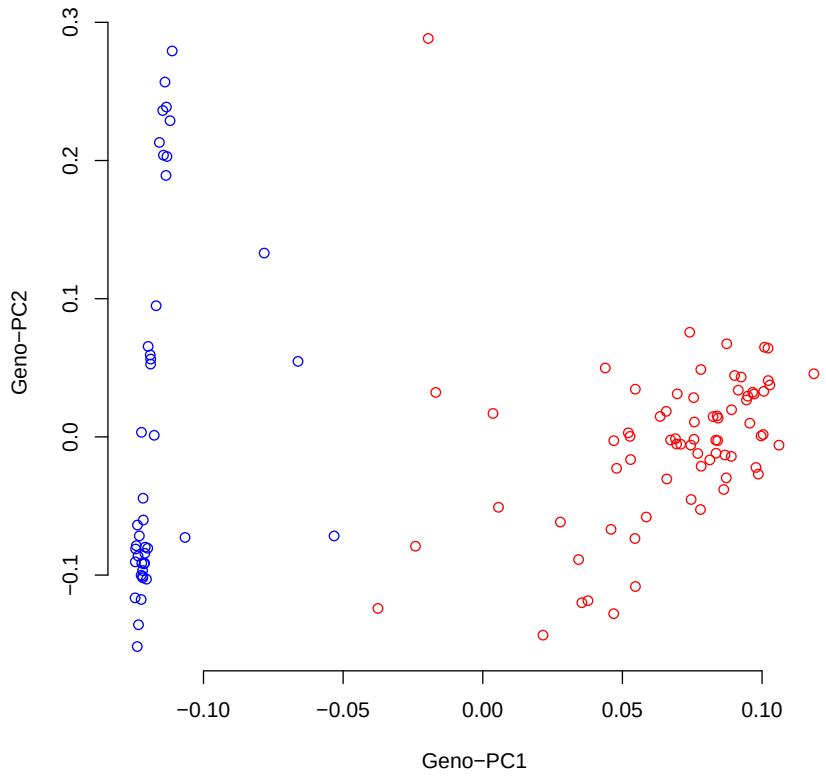

Figure 5.7: The Principal Component Plot for the genotype showing the population structure of the data, where the first principal component is plotted against the second, with red color indicate African Americans and blue indicate Europeans.

BIBLIOGRAPHY

[1] S. Schuster, Nature Methods **5**, 16 (2007), ISSN 1548-7091.

[2] M. Metzker, Nature reviews. Genetics **11**, 31 (2010), ISSN 1471-0064.

[3] M. DePristo, E. Banks, *et al.*, Nat Genet **43**, 491 (2011), ISSN 1061-4036.

[4] T. I. H. . Consortium, Nature **467**, 52 (2010), ISSN 0028-0836.

[5] L. Hindorff, P. Sethupathy, *et al.*, Proceedings of the National Academy of Sciences **106**, 9362 (2009), ISSN 1091-6490.

[6] Z. Wang, M. Gerstein, *et al.*, Nat Rev Genet **10**, 57 (2009), ISSN 1471-0064.

[7] W. Cookson, L. Liang, *et al.*, Nature reviews. Genetics **10**, 184 (2009), ISSN 1471-0064.

[8] A. Torkamani, E. Topol, *et al.*, Genomics **92**, 265 (2008), ISSN 1089-8646.

[9] Y.-A. Kim, S. Wuchty, *et al.*, PLoS computational biology **7**, e1001095 (2011), ISSN 1553-7358.

[10] J. Zhu, B. Zhang, *et al.*, Nature Genetics **40**, 854 (2008), ISSN 1061-4036.

[11] X. Yang, J. Deignan, *et al.*, Nature Genetics **41**, 415 (2009), ISSN 1061-4036.

[12] T. Manolio, F. Collins, *et al.*, Nature **461**, 747 (2009), ISSN 0028-0836.

[13] P. Vineis and N. Pearce, Nature Reviews Genetics **11**, 589 (2010), ISSN 1471-0056.

[14] S. Lee, N. Wray, *et al.*, Am J Hum Genet **88**, 294 (2011).

[15] D. Reich and D. Goldstein, Genetic epidemiology **20**, 4 (2001), ISSN 0741-0395.

[16] J. Pritchard, M. Stephens, *et al.*, The American Journal of Human Genetics **67**, 170 (2000).

[17] J. Pritchard, M. Stephens, *et al.*, Genetics **155**, 945 (2000), ISSN 0016-6731.

[18] A. Price, N. Patterson, *et al.*, Nature Genetics **38**, 904 (2006), ISSN 1061-4036.

[19] E. Setakis, H. Stirnadel, *et al.*, Genome Res. **16**, 290 (2006).

[20] H. M. Kang, N. Zaitlen, *et al.*, Genetics **178**, 1709 (2008), ISSN 0016-6731.

[21] B. G. Harvey, A. Heguy, *et al.*, J Mol Med (Berl) **85**, 39 (2007).

[22] M. Corbex, O. Poirier, *et al.*, Genet Epidemiol **19**, 64 (2000).

[23] H. Miyazaki, N. Oka, *et al.*, Hypertens Res **29**, 1029 (2006), URL `http://www.hubmed.org/fulltext.cgi?uids=17378376`.

[24] J. Listgarten, C. Kadie, *et al.*, Proceedings of the National Academy of Sciences of the United States of America **107**, 16465 (2010), ISSN 1091-6490.

[25] O. Stegle, L. Parts, *et al.*, PLoS computational biology **6**, e1000770 (2010), ISSN 1553-7358.

[26] C. Friguet, M. Kloareg, *et al.*, Journal of the American Statistical Association **104**, 1406 (2009), ISSN 0162-1459.

[27] J. Leek and J. Storey, PLoS Genet **3**, e161 (2007), ISSN 1553-7404.

[28] L. Parts, O. Stegle, *et al.*, PLoS Genet **7**, e1001276 (2011), ISSN 1553-7404.

[29] J. Leek, E. Johnson, *et al.*, Bioinformatics **28**, 882 (2012), ISSN 1460-2059.

[30] O. Stegle, L. Parts, *et al.*, Nat Protoc **7**, 500 (2012), URL `http://www.hubmed.org/fulltext.cgi?uids=22343431`.

[31] N. Fusi, O. Stegle, *et al.*, PLoS Comput Biol **8** (2012), URL http://www.hubmed.org/fulltext.cgi?uids=22241974.

[32] H. M. Kang, C. Ye, *et al.*, Genetics **180**, 1909 (2008), ISSN 0016-6731.

[33] D. Rubin and D. Thayer, Psychometrika **47**, 69 (1982).

[34] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)* (Chapman and Hall/CRC, 1988), 1st ed., ISBN 0412014211, URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0412014211.

[35] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, 2007), 1st ed., ISBN 0387310738.

[36] V. Emilsson, G. Thorleifsson, *et al.*, Nature **452**, 423 (2008), ISSN 0028-0836.

[37] E. Schadt, S. Monks, *et al.*, Nature **422**, 297 (2003), ISSN 0028-0836.

[38] B. Stranger, M. Forrest, *et al.*, PLoS Genetics **1**, e78 (2005).

[39] A. Dixon, L. Liang, *et al.*, Nature Genetics **39**, 1202 (2007), ISSN 1061-4036.

[40] R. Spielman, L. Bastone, *et al.*, Nature genetics **39**, 226 (2007), ISSN 1061-4036.

[41] V. Cheung and R. Spielman, Nature reviews. Genetics **10**, 595 (2009), ISSN 1471-0064.

[42] S. Montgomery and E. Dermitzakis, Human molecular genetics **18**, R211 (2009), ISSN 1460-2083.

[43] A. Brunner, D. Johnson, *et al.*, Genome research **19**, 1044 (2009), ISSN 1088-9051.

[44] G. Robertson, M. Hirst, *et al.*, Nat Meth **4**, 651 (2007), ISSN 1548-7091.

[45] J. Pickrell, J. Marioni, *et al.*, Nature **464**, 768 (2010), ISSN 0028-0836.

[46] D. Arends, J. van der Velde, *et al.*, Bioinformatics (Oxford, England) (2012), ISSN 1367-4811.

[47] A. Nica and E. Dermitzakis, Human Molecular Genetics **17**, R129 (2008), ISSN 1460-2083.

[48] E. Schadt, J. Lamb, *et al.*, Nature genetics **37**, 710 (2005), ISSN 1061-4036.

[49] E. Dermitzakis, Nature Genetics **40**, 492 (2008), ISSN 1061-4036.

[50] C. Damerval, A. Maurice, *et al.*, Genetics **137**, 289 (1994), URL `http://www.hubmed.org/fulltext.cgi?uids=7914503`.

[51] J. Michaelson, S. Loguercio, *et al.*, Methods **48**, 265 (2009), ISSN 10462023.

[52] Harvey, Ben-Gary, *et al.*, Journal of Molecular Medicine **86**, 853 (2008), ISSN 0946-2716.

[53] G. Gibson, Nature reviews. Genetics **9**, 575 (2008), ISSN 1471-0064.

[54] H. M. Kang, J. H. Sul, *et al.*, Nat Genet **42**, 348 (2010), URL `http://www.hubmed.org/fulltext.cgi?uids=20208533`.

[55] Q. Shi, Z. Zhang, *et al.*, Pharmacogenet Genomics **15**, 547 (2005), URL `http://www.hubmed.org/fulltext.cgi?uids=16006998`.

[56] W. Wu, H. Liu, *et al.*, Lung Cancer **63**, 180 (2009), ISSN 0169-5002, URL http://www.sciencedirect.com/science/article/pii/S0169500208002456.

[57] E. Dehan, A. Ben-Dor, *et al.*, Lung cancer (Amsterdam, Netherlands) **56**, 175 (2007), ISSN 0169-5002.

[58] M. Minczuk, J. He, *et al.*, Nucleic Acids Research **39**, 4284 (2011), http://nar.oxfordjournals.org/content/39/10/4284.full.pdf+html, URL http://nar.oxfordjournals.org/content/39/10/4284.abstract.

[59] B. Devlin and K. Roeder, Biometrics **55**, 997 (1999), ISSN 0006-341X.

[60] G. Chen, P. Marjoram, *et al.*, Genome research **19**, 136 (2009), ISSN 1088-9051.

[61] B. Voight, A. Adams, *et al.*, PNAS **102**, 18508 (2005).

[62] T. Raman, T. O'Connor, *et al.*, BMC Genomics **10**, 493 (2009), ISSN 1471-2164.

[63] M. Dai, P. Wang, *et al.*, Nucleic Acids Research **33**, e175 (2005), ISSN 1362-4962.

[64] R. Irizarry, B. Hobbs, *et al.*, Biostatistics **4**, 249 (2003), ISSN 1468-4357.

[65] R. Irizarry, B. Bolstad, *et al.*, Nucleic acids research **31**, e15 (2003), ISSN 1362-4962.

[66] J. Wigginton, D. Cutler, *et al.*, American journal of human genetics **76**, 887 (2005), ISSN 0002-9297.

[67] S. Purcell, B. Neale, *et al.*, American journal of human genetics **81**, 559 (2007), ISSN 0002-9297.

[68] R. Edgar, M. Domrachev, *et al.*, Nucleic Acids Research **30**, 207 (2002), ISSN 1362-4962.

[69] B. Engelhardt and M. Stephens, PLoS Genetics **6**, e1001117 (2010), ISSN 1553-7404.

[70] B. Devlin, S. Bacanu, *et al.*, Nat Genet **36** (2004), ISSN 1061-4036.

[71] Y. Aulchenko, S. Ripke, *et al.*, Bioinformatics (Oxford, England) **23**, 1294 (2007), ISSN 1367-4811.

[72] W. Yu, A. Wulf, *et al.*, European Journal of Human Genetics **aop** (????).

[73] T. E. Thorgeirsson, F. Geller, *et al.*, Nature **452**, 638 (2008).

[74] A. Caspi, K. Sugden, *et al.*, Science **301**, 386 (2003).

[75] S. R. Kleeberger and D. Peden, Annu Rev Med **56**, 383 (2005).

[76] M. J. Khoury and S. Wacholder, Am J Epidemiol **169**, 227 (2009).

[77] T. Yoshida and R. M. Tuder, Physiol Rev **87**, 1047 (2007).

[78] S. S. Hecht, Nat Rev Cancer **3**, 733 (2003).

[79] W. MacNee, Eur J Pharmacol **429**, 195 (2001).

[80] L. Mucha, J. Stephenson, *et al.*, Gend Med **3**, 279 (2006).

[81] J. M. Sethi and C. L. Rochester, Clin Chest Med **21**, 67 (2000).

[82] E. Puchelle, J. M. Zahm, *et al.*, Proc Am Thorac Soc **3**, 726 (2006).

[83] P. Maestrelli, M. Saetta, *et al.*, Am J Respir Crit Care Med **164**, 76 (2001).

[84] Z. Ammous, N. R. Hackett, *et al.*, Chest **133**, 1344 (2008).

[85] K. Steiling, A. Y. Kadar, *et al.*, PLoS One **4** (2009).

[86] S. Pierrou, P. Broberg, *et al.*, Am J Respir Crit Care Med **175**, 577 (2007).

[87] A. Spira, J. E. Beane, *et al.*, Nat Med **13**, 361 (2007).

[88] K. Juul, A. Tybjaerg-Hansen, *et al.*, Am J Respir Crit Care Med **173**, 858 (2006).

[89] H. Takizawa, M. Tanaka, *et al.*, Am J Respir Crit Care Med **163**, 1476 (2001).

[90] W. Cookson, L. Liang, *et al.*, Nat Rev Genet **10**, 184 (2009).

[91] Y. Benjamini and Y. Hochberg, Journal of the Royal Statistical Society. Series B (Methodological) **57**, 289 (1995), ISSN 00359246.

[92] W. He, Q. Liu, *et al.*, Mol Immunol **44**, 2850 (2007), URL `http://www.hubmed.org/fulltext.cgi?uids=17328955`.

[93] X. Wu, K. Hepner, *et al.*, Proc Natl Acad Sci U S A **97**, 4233 (2000), URL `http://www.hubmed.org/fulltext.cgi?uids=10760291`.

[94] X. Cao, G. Wei, *et al.*, J Immunol **172**, 4851 (2004), URL `http://www.hubmed.org/fulltext.cgi?uids=15067063`.

[95] H. Suzuki, M. Ouchida, *et al.*, Lung Cancer **59**, 24 (2008), URL `http://www.hubmed.org/fulltext.cgi?uids=17854949`.

[96] M. Sato, Y. Mori, *et al.*, Hum Genet **103**, 96 (1998), URL `http://www.hubmed.org/fulltext.cgi?uids=9737784`.

[97] Centers for Disease Control and Prevention (CDC), MMWR Morb Mortal Wkly Rep **57**, 1221 (2008).